**MASTER EN SCIENCES MATHÉMATIQUES**

# Classification pyramidale

Michel, Aline

*Award date:*
2004

FUNDP
Faculté des Sciences
Département de Mathématique

Rempart de la Vierge, 8
B-5000 Namur Belgique

# Classification pyramidale

Mémoire présenté pour l'obtention
du grade de
Licencié en Sciences Mathématiques
par

Aline Michel

**Promoteur** : André Hardy

Année Académique 2003-2004

*Je remercie spécialement le Professeur André Hardy ainsi que l'Assistante Pascale Lallemand pour leur disponibilité et les précieux conseils apportés tout au long de l'élaboration de ce travail.*
*Je tiens également à remercier ma famille, Jordan, ses parents et mes amies pour leur présence et leur soutien durant ces quatre années d'étude.*

# Résumé

Le but de ce travail est d'étudier la classification pyramidale de données classiques et symboliques. Nous commençons par introduire les différents types de variables ainsi que les notions de compatibilité entre un ordre et un indice nécessaires à l'introduction des pyramides. Ensuite, nous détaillons la visualisation des pyramides en termes de structure pyramidale. Les concepts et propriétés d'indiçage et d'indice pyramidal sont alors présentés. L'ensemble des pyramides étant une extension de l'ensemble des hiérarchies, les liens entre ces deux familles sont ensuite développés afin de résoudre le problème des pyramides saturées. Finalement, nous donnons un algorithme de classification pyramidale fourni par le logiciel SODAS 2 et nous terminons par l'étude d'applications de la classification pyramidale à des ensembles de données artificielles et réelles symboliques.

# Abstract

The aim of this work is to study the pyramidal clustering of classic and symbolic data. We begin by introduce the different types of variables as well as the notions of compatibility between an order and an index which are necessary to the introduction of pyramids. Next we detail the visualization of pyramids in terms of pyramidal structure. Concepts and properties of indexing and pyramidal index are then presented. Since the set of pyramids is an extension of the set of hierarchies, ties between this two families are developed in order to resolve the problem of saturated pyramids. Finally we give an algorithm of pyramidal clustering provided by the software SODAS 2 and we end by applications of the pyramidal clustering to artificial and real symbolic data.

# Table des matières

# Introduction

Depuis toujours, l'intérêt de classer des éléments réels existe et pour trouver des classes, la classification hiérarchique est un méthode connue depuis bien longtemps, où une hiérarchie est formée d'une suite de partitions emboîtées. Cependant, la complexité de la réalité a fait que un autre moyen de représentation des classes s'est développé, les classes empiétantes. La classification pyramidale fourni justement ce type de classes en les représentant par des recouvrements emboîtés.

Ce travail est séparé en deux parties. Une première partie développe tous les aspects théoriques servant à introduire et à décrire les pyramides. Et une seconde partie pratique présente un programme de classification hiérarchique et pyramidale suivi d'applications pratiques.

Nous commencons la partie théorique en décrivant les différents types de variables, classiques et symboliques, servant à représenter les éléments à classer. Ces variables permettent deux types de classification, la classification classique et symbolique. Ensuite, nous établissons les différentes notions indispensables pour introduire les pyramides, notamment celles de compatibilité. Après, nous énonçons certaines propriétés concernant les hiérarchies et les ultramétriques qui serviront à montrer, plus loin, que le modèle pyramidal étend bien le modèle hiérarchique, c'est-à-dire que les hiérarchies sont en fait des pyramides particulières, et que l'ensemble des ultramétriques est inclus dans l'ensemble des indices pyramidaux. Enfin, nous donnons la définition exacte d'une pyramide et nous détaillons également les différents aspects de la représentation visuelle des pyramides. Ensuite, nous introduisons le concept d'indiçage d'une pyramide qui permet d'associer une hauteur à chaque palier d'une pyramide et nous donnons aussi la définition d'indice pyramidal. Dès lors, nous pouvons démontrer un résultat fondamental selon lequel il existe une bijection entre l'ensemble des indices pyramidaux et

l'ensemble des pyramides indicées "au sens large". Au niveau de la représentation visuelle, une pyramide peut devenir difficile à interpréter lorsque ses paliers sont trop nombreux. Afin de réduire le nombre de paliers, on profite du fait que les pyramides constituent une extension des hiérarchies pour proposer une méthode de "hiérarchisation" d'une pyramide et une méthode de "pyramidisation" d'une hiérarchie servant à faire apparaître des classes recouvrantes. On donne également une technique d'épuration, c'est-à-dire de suppression d'arêtes inutiles.

D'un point de vue plus pratique, nous abordons ensuite, dans la deuxième partie, un programme, appelé HIPYR, de classification hiérarchique et pyramidale permettant de classer soit des données classiques, soit des données symboliques et nous décrivons son fonctionnement et son algorithme. Finalement, nous terminons en illustrant cette théorie par quelques applications concernant, en particulier, la classification pyramidale symbolique.

# Première partie

# Cadre Théorique

# Chapitre 1

# Les données classiques et symboliques

## 1.1 Les données classiques

Pour définir les variables classiques considérons
- $\Omega = 1, \cdots, n$ : l'ensemble des $n$ individus
- $Y_1, \cdots, Y_p$ : les $p$ variables mesurées sur chaque individu
- $\mathcal{Y}_1, \cdots, \mathcal{Y}_p$ : les espaces d'observation associés à ces $p$ variables, c'est à dire par exemple que $\mathcal{Y}_j$ est l'ensemble des valeurs pouvant être prises par la variable $Y_j$ $(j \in 1, \cdots, p)$

Formellement, une **variable classique** est définie de la façon suivante :

$$\begin{array}{rccc} Y_j : & \Omega & \rightarrow & \mathcal{Y}_j \\ & k & \mapsto & Y_j(k) = x_{kj} \end{array}$$

Toutes ces valeurs sont placées dans une matrice de données :

$$\tilde{X} = (x_{kj})_{n \times p}$$

On distingue deux types de variables classiques en fonction de la taille de leurs espaces d'observations et de la structure imposée aux éléments de ces espaces :

1. les variables quantitatives et
2. les variables qualitatives,

elles sont décrites dans les deux points suivants.

### 1.1.1 Les variables quantitatives

Une variable $Y$ est dite **quantitative** si son espace d'observation $\mathcal{Y}$ est tel que $\mathcal{Y} \subseteq \mathbb{R}$.

#### Variable quantitative continue

Une variable quantitative $Y$ est **continue** si elle peut prendre un nombre infini non dénombrable de valeurs dans $\mathbb{R}$.

Dés lors, les 3 situations possibles sont :
$\mathcal{Y} = \mathbb{R}$,
$\mathcal{Y} = \mathbb{R}^+$ ou
$\mathcal{Y} = [a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$.

**Exemple 1.1.1** *La variable $Y$ = "Le chiffre d'affaires d'un restaurant en milliers d'euros", admet l'espace d'observation $\mathcal{Y} = \mathbb{R}^+$.*

**Exemple 1.1.2** *La variable $Y$ = "Le poids d'une personne adulte en kilos" admet l'espace d'observation $\mathcal{Y} = [a, b] = [30, 250] \in \mathbb{R}$.*

#### Variable quantitative discrète

Une variable quantitative $Y$ est **discrète** si son espace d'observation $\mathcal{Y}$ contient un nombre fini ou un nombre infini dénombrable de valeurs $\xi_i \in \mathbb{R}$.

Donc, soit $\mathcal{Y} = \{\xi_1, \ldots, \xi_N\} \subset \mathbb{R}$, soit $\mathcal{Y} = \{\xi_1, \xi_2, \ldots\} \subset \mathbb{R}$.

**Exemple 1.1.3** *La variable $Y$ = "le nombre de portes d'un immeuble" admet $\mathcal{Y} = \{1, 2, \ldots, N\}$ où $N < \infty$.*

**Exemple 1.1.4** *La variable $Y$ = "le nombre d'accidents de la route dans un pays" (au cours d'une année donnée) admet $\mathcal{Y} = \mathbb{N}_0 := \{0, 1, 2, \ldots\}$, l'ensemble des entiers y compris 0.*

### 1.1.2 Les variables qualitatives

Une variable $Y$ est dite **qualitative** (ou **catégorique**) si le nombre de valeurs de l'espace d'observation $\mathcal{Y}$ est fini et si les éléments de $\mathcal{Y}$, appelés catégories ici, ne portent aucune structure numérique (mais peut être une autre structure).

#### Variable nominale

Une variable qualitative $Y$ est dite **nominale** si elle a des modalités distinctes les unes des autres, mais sans structure interne, c'est-à-dire, sans possibilité d'ordre ou de calcul entre elles.

Dans ce cas, pour deux catégories $x, y \in \mathcal{Y}$, nous ne pouvons distinguer entre elles que deux alternatives :

$$x = y \text{ ou } x \neq y.$$

Deux individus $k, l \in \Omega$ peuvent par conséquent être égaux ou inégaux sur cette variable mais aucune autre distinction n'est possible.
La mesure de proximité $\delta(x, y)$ entre deux catégories $x, y \in \mathcal{Y}$ se définit par :

$$\delta(x, y) = \begin{cases} 1 & \text{si x=y,} \\ 0 & \text{sinon.} \end{cases}$$

**Exemple 1.1.5** *(- tiré de [3]) La variable $Y$ = "état civil" a pour espace d'observation $\mathcal{Y} = \{$marié, veuf, divorcé, célibataire$\}$.*

Dans le cas particulier où l'espace d'observation $\mathcal{Y}$ ne comprend que deux alternatives, codées habituellement par 0 et 1, on parle de variables **binaires** ou **dichotomiques**. ($\mathcal{Y} = \{0, 1\}$ avec $|\mathcal{Y}| = 2$)

**Exemple 1.1.6** *(- tiré de [3]) La variable $Y$ = "Sexe" a comme espace d'observation $\mathcal{Y} = \{$Féminin, Masculin$\}$ que nous pouvons coder par $0, 1$.*

Pour une commodité de notation, lorsque la variable nominale considérée a plus de deux modalités, les $s$ catégories peuvent aussi être codées par $0, 1, 2, \ldots, s-1$. Cependant, aucune opération arithmétique avec ces codes ne peut être définie, ni même de mesure de proximité.

**Exemple 1.1.7** *(- tiré de [3]) Les catégories de l'espace d'observation correspondant à la variable $Y$ = "état civil" peuvent être codées : $\mathcal{Y} = \{0, 1, 2, 3\}$ où les codes 0, 1, 2 et 3 correspondent, respectivement, aux modalités "marié", "veuf", "divorcé" et "célibataire".*

### Variable ordinale

Une variable qualitative $Y$ est appelée **ordinale** si elle a des modalités pouvant être hiérarchisées entre elles, aucun calcul ne peut cependant être défini.

Cela revient à dire que l'espace d'observation $\mathcal{Y}$ est muni d'un ordre linéaire total $\prec$ tel que $\forall x, y \in \mathcal{Y}$, $x \neq y$, nous avons que $x \prec y$ ou $y \prec x$.

Tout comme pour les variables nominales, les différentes catégories peuvent être codées de sorte que $\mathcal{Y} = \{0, 1, \ldots, s-1\}$ où $s$ est le nombre de modalités pouvant être prises par cette variable. Il est bien entendu proscrit de définir des opérations arithmétiques avec ces codes. Cependant, ils peuvent servir à la définition d'une mesure de proximité par :

$$\delta(x, y) = |x - y| \qquad \forall x, y \in \mathcal{Y}$$

qui représente en fait le nombre de catégories de $\mathcal{Y}$ strictement comprises entre $x$ et $y$ selon l'ordre total imposé par $\prec$.

**Exemple 1.1.8** *La variable $Y$ = "Popularité d'un chanteur" a pour espace d'observation $\mathcal{Y} = \{$détesté, supporté, ..., populaire, très apprécié$\}$.*

## 1.1.3 Vecteurs et matrice de données

Comme précédemment, considérons l'ensemble des $n$ individus $\Omega = \{1, 2, \ldots, n\}$ caractérisés par $p$ variables notées $Y_1, \ldots, Y_p$, $\mathcal{Y}_j$ représentant l'espace d'observation de la variable $Y_j$ où $j = 1, \ldots, p$.
Nous désignons par $X$ le vecteur colonne à $p$ dimensions des $p$ variables $Y_1, \ldots, Y_p$ :

$$X = \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix} \in \mathcal{X} = \bigotimes_{i=1}^{p} \mathcal{Y}_j$$

où $\bigotimes_{i=1}^{p} \mathcal{Y}_j$ est le produit cartésien des $p$ espaces d'observation $\mathcal{Y}_1, \ldots, \mathcal{Y}_p$.

La valeur ou la catégorie de $Y_j$ qui est observée pour l'individu $k \in \Omega$ est notée $x_{kj} = Y_j(k)$.

Pour chaque individu $k \in \Omega$, les $p$ observations $x_{k1}, \ldots, x_{kp}$ sont représentées dans un vecteur colonne à $p$ dimensions

$$x_k = X(k) = \begin{pmatrix} x_{k1} \\ x_{k2} \\ \vdots \\ x_{kp} \end{pmatrix} \in \mathcal{X} = \bigotimes_{i=1}^{p} \mathcal{Y}_j \ .$$

En prenant en compte toutes les $np$ données ensemble, nous obtenons la matrice de données classiques qui a la forme suivante :

$$\tilde{X} = (x_{kj})_{n \times p} = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} = (y_1, \ldots, y_p)$$

tel que la $k$-ième ligne $x_k'$ contient les données observées pour l'individu $k$ et la $j$-ième colonne $y_j$ représente les valeurs prises par la variable $Y_j$ pour les $n$ individus.

### 1.1.4 Exemple

Soit $\Omega = \{$Luc, Bernadette, Charles, Jacques, Françoise$\}$, un ensemble de cinq personnes pour lesquelles quatre variables ont été considérées :

- $Y_1$ : la taille en cm (variable quantitative continue)
- $Y_2$ : le sexe (0 =femme ou 1 =homme) (variable qualitative nominale binaire)
- $Y_3$ : le grade obtenu pour une session d'examens (Ajournement, Satisfaction, Distinction, Grande Distinction, la Plus Grande Distinction) (variable qualitative ordinale)
- $Y_4$ : la nationalité (variable qualitative nominale)

La matrice de données $\tilde{X}$ $(5 \times 4)$ correspondante est de la forme suivante :

| | $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ |
|---|---|---|---|---|
| Luc | 165 | 0 | S | Belge |
| Bernadette | 180 | 1 | D | Française |
| Charles | 159 | 0 | GD | Grecque |
| Jacques | 175 | 0 | PGD | Français |
| Françoise | 170 | 1 | GD | Belge |

## 1.2 Les données symboliques

Parmi les données symboliques, nous distinguons trois types de variables :
- les variables multivaluées,
- les variables de type intervalle et
- les variables modales.

L'ensemble des objets $E$ peut être défini de deux façons différentes :

1. un ensemble $E = \Omega = \{1, \ldots, n\}$ d'individus appelés **objets du premier ordre.**
2. un ensemble $E = \{C_1, C_2, \ldots\}$ de classes $C_i \subseteq \Omega$ d'individus appelées **objets du second ordre.**

### 1.2.1 Variables multivaluées

La variable $Y$, dont l'espace d'observation est $\mathcal{Y}$, est dite **à valeurs dans un ensemble** $\mathcal{B}$ lorsque :

$$\begin{array}{rccll} Y: & E & \rightarrow & \mathcal{B} & \forall k \in E \\ & k & \mapsto & Y(k) & \end{array}$$

où $\mathcal{B} = \mathcal{P}(\mathcal{Y}) = \{U \neq \emptyset \mid U \subseteq \mathcal{Y}\}$.

Une variable $Y$ est dite **multivaluée** lorsque les valeurs $Y(k)$ sont toutes des sous-ensembles finis de $\mathcal{Y}$, c'est-à-dire

$$|Y(k)| < \infty, \forall k \in E.$$

Une variable $Y$ est dite **multivaluée catégorique** si $\mathcal{Y}$ a un nombre fini de catégories et donc

$$|Y(k)| < \infty, \forall k \in E.$$

Une variable est dite **multivaluée quantitative** si les valeurs $Y(k)$ sont des ensembles finis de nombres réels, c'est-à-dire :

$$Y(k) \subset \mathbb{R} \text{ et } |Y(k)| < \infty, \forall k \in E.$$

**Exemple 1.2.1** *(Variable multivaluée - tiré de [3])*
Considérons :

- $E$ = {Bruxelles, Charleroi, Liège, Namur}
- $Y_1$ = les salles de cinéma présentes dans une ville
  $\mathcal{Y}_1$ = {Caméo, Carollywood, Eldorado, Forum, Kinépolis, Le Parc, UGC, ...}
- $Y_2$ =le chiffre d'affaires des deux plus grands cinémas de la ville en euros
  $\mathcal{Y}_2 = \mathbb{R}^+$

Voici les résultats pour les 5 villes de l'ensemble $E$ :

| | $Y_1$ | $Y_2$ |
|---|---|---|
| Bruxelles | {Aventure, Kinépolis, Movy Club, Styx, UGC} | {22500, 18750} |
| Charleroi | {Carollywood, Le Parc, Paradiso} | {5000, 3750} |
| Liège | {Kinépolis, Le Parc, Opéra, Palace, UGC} | {21250, {20250} |
| Namur | {Acinapolis, Caméo, Eldorado, Forum} | {8750, 7500} |

Dans cet exemple,
$Y_1$ est une variable multivaluée catégorique et
$Y_2$ est une variable multivaluée quantitative.

### 1.2.2 Variables de type intervalle

Une variable est dite de type **intervalle** si $\forall\ k \in E$, l'ensemble $Y(k)$ est un intervalle borné et fermé de $\mathbb{R}$.

Dans ce cas $\mathcal{B} = \mathcal{I}$, c'est-à-dire que $\mathcal{B}$ est l'ensemble des intervalles fermés bornés de $\mathbb{R}$.

**Exemple 1.2.2** *(Variable de type intervalle)*
Soient :

- $E = \{$ femmes au foyer d'un village$\}$
- $Y =$ le temps passé quotidiennement à faire les tâches ménagères (en heures)
- $\mathcal{Y} = \{[a, b] \mid a, b \in \mathbb{R}^+, 0 \leq a \leq b \leq 24 < \infty\}$

On peut avoir les résultats suivants :

$$Y(k) = [0, 2],\ Y(l) = [2, 4], \ldots \text{ où } k, l \in E.$$

### 1.2.3 Variables multivaluées et intervalles par agrégation

Supposons que :

- $\Omega = \{1, \ldots, n\}$ est l'ensemble des objets du premier ordre,
- $\widetilde{Y}$ est une variable univaluée classique et
- $E = \{C_1, \ldots, C_m\}$ est l'ensemble des classes $C_i \subseteq \Omega$, appelées objets du second ordre.

Nous cherchons à caractériser le comportement de ces classes par rapport à la variable $\widetilde{Y}$. Une solution est de définir une variable "globale" ou "agrégée" $Y$ qui spécifie les valeurs prises par $\widetilde{Y}$ sur les classes $C_i$.

**Exemple 1.2.3** *(Variable multivaluée obtenue par agrégation)*
Soient :

- $\Omega = \{$membres d'un club tennis$\}$
- $E = \{C_1, \ldots, C_m\} = \{m$ catégories d'âge$\}$
- $\widetilde{Y}(k) =$ les points gagnés à un tournoi de tennis par la personne $k$

La description de la catégorie d'âge $C_i$ sera donnée par

$$Y(C_i) = \{50, 75, 100, 125, 150, 175, 200\},$$

c'est-à-dire l'ensemble des meilleures performances sur 100 mètres des coureurs de la catégorie d'âge $C_i$.

**Exemple 1.2.4** *(Variable intervalle obtenue par agrégation)*
Considérons $\Omega$, $E$ et $\widetilde{Y}$ de l'exemple précédent. Nous pouvons décrire la

catégorie d'âge $C_i$ par $Y(C_i) = [\alpha, \beta]$ où

$$\alpha = \min_{\omega \in C_i} \{\widetilde{Y}(\omega)\}$$

$$\beta = \max_{\omega \in C_i} \{\widetilde{Y}(\omega)\}$$

et par conséquent, $Y(C_i) = [50, 200]$.

### 1.2.4 Variables modales

Une variable **modale** $Y$ d'espace d'observation $\mathcal{Y}$ sur $E = \Omega = \{1, 2, \ldots, n\}$ est une variable multivaluée pour laquelle :

- $\forall k \in E$, $Y(k) \subset \mathcal{Y}$
- $\forall y \in Y(k)$, nous définissons un poids, une probabilité ou une fréquence $w(y)$ qui indique la pertinence de la catégorie $y$ pour l'objet $k$.

Plus formellement, une variable **modale** $Y$ sur un ensemble $E = \Omega = \{1, 2 \ldots, n\}$ d'objets à valeurs dans $\mathcal{Y}$ est une fonction définie par :

$$Y(k) = (U(k), \pi_k), \ \forall k \in E$$

où

- $\pi_k$ est une mesure ou une distribution (poids, probabilité ou fréquence) sur les valeurs possibles de $\mathcal{Y}$ et
- $U(k) \subseteq \mathcal{Y}$ est le support de $\pi_k$ dans le domaine $\mathcal{Y}$.

**Exemple 1.2.5** *(Variable modale)*
Soient :

- $\Omega = \{1, \ldots, 100\}$ un ensemble de 100 fonctionnaires,
- $\widetilde{Y}$ une variable univaluée classique qui mesure la taille de ces fonctionnaires en centimètres et
- $C = \{1, \ldots, 10\}$ la classe des 10 premiers fonctionnaires.

$\widetilde{Y}(C)$ prend les valeurs suivantes : 167, 158, 186, 174, 168, 182, 170, 154, 179, 172. La variable modale $Y$ qui décrit la taille dans la classe $C$ peut avoir une réalisation sous la forme d'un histogramme :

$$Y(C) = \left\{ (]150, 160], \frac{2}{10}), (]160, 170]), \frac{3}{10}), (]170, 180], \frac{3}{10}), (]180, 190], \frac{2}{10}) \right\}$$

### 1.2.5 Résumé des types de données symboliques

Une **variable symbolique** d'espace d'observation $\mathcal{Y}$ est définie de la manière suivante :

$$\begin{array}{rccll} Y: & E & \rightarrow & \mathcal{B} & \forall k \in E \\ & k & \mapsto & Y(k) & \end{array}$$

$$\text{où } \mathcal{B} = \mathcal{P}(\mathcal{Y}) = \{U \neq \emptyset \mid U \subseteq \mathcal{Y}\} .$$

1. Si $\mathcal{B} = \mathcal{Y}$, nous sommes dans le cas d'une variable classique univaluée.
2. $Y$ est à valeurs dans un ensemble $\mathcal{B}$ si $Y(k) \subseteq \mathcal{Y}$ $\forall k \in E$, ce qui revient à considérer $\mathcal{B} = \mathcal{P}(\mathcal{Y})$.
3. Variable intervalle : $Y$ est une variable de type intervalle si, $\forall k \in E$, $Y(k) = [\alpha, \beta]$ est un intervalle de $\mathcal{Y}$ et donc $\mathcal{B}$ est l'ensemble $\mathcal{J}$ des intervalles fermés bornés dans $\mathcal{Y}$.
4. Variable multivaluée : $Y$ est une variable multivaluée (catégorique ou quantitative) si $Y(k) \subseteq \mathcal{Y}$ et $|Y(k)| < \infty$, $\forall k \in E$.
5. Variable modale : $Y$ est une variable modale d'espace d'observation $\mathcal{Y}$ si, $\forall k \in E$, $Y(k) = \pi_a$ est une mesure non-négative sur $\mathcal{Y}$, habituellement une distribution de fréquence, de probabilité ou un poids, d'où $\mathcal{B} = \mathcal{M}(Y)$.

Ceci nous permet de considérer les données symboliques comme une extension des données classiques.

### 1.2.6 Le tableau des données symboliques

Considérons un ensemble de base $E = \{1, \ldots, N\}$ où les éléments $u \in E$ sont appelés des objets de $E$.

$E$ peut être décrit de différentes manières :
- $E = \Omega = \{1, \ldots, n\}$ un ensemble de $n$ individus $(N = n)$ ou
- $E \subset \Omega$ un échantillon de $\Omega$, $(N \leq n)$ ou
- un ensemble $E = \{C_1, \ldots, C_m\}$ de classes $C_1, \ldots, C_m \subseteq \Omega$ d'individus $k \in \Omega$ ou objets du second ordre $(N = m)$.

Supposons que les propriétés de chaque objet $u \in E$ soient décrites par $p$ variables symboliques $Y_1, \ldots, Y_p$ où $Y_j$ a pour espace d'observation $\mathcal{Y}_j$.

Notons par $X(u) = (Y_1(u), \ldots, Y_p(u))'$ le vecteur des variables symboliques déterminées pour $u \in E$. Ainsi, chaque objet $u \in E$ peut être décrit par un vecteur de données symboliques :

$$x_u = X(u) = \begin{pmatrix} \xi_{u1} \\ \vdots \\ \xi_{up} \end{pmatrix} \in \mathcal{X} = \bigotimes_{i=1}^{p} \mathcal{B}_j$$

où

- $\xi_{uj} = Y_j(u) \in \mathcal{B}_j$ est la valeur de la j-ième variable symbolique $Y_j$ pour l'individu $u$ $(j = 1, \ldots, p)$ ;
- $\bigotimes_{i=1}^{p} \mathcal{B}_j$ est le produit cartésien des $p$ espaces d'observation $\mathcal{B}_1, \ldots, \mathcal{B}_p$.

Toutes ces données peuvent être compilées dans une **matrice de données symboliques** :

$$\underline{X} = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_N' \end{pmatrix} = \begin{pmatrix} \xi_{11} & \xi_{12} & \cdots & \xi_{1p} \\ \xi_{21} & \xi_{22} & \cdots & \xi_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{N1} & \xi_{N2} & \cdots & \xi_{Np} \end{pmatrix} = (\xi_{uj})_{N \times p}$$

Ici, une cellule $\xi_{uj}$ de la matrice de données symboliques peut contenir des ensembles, des intervalles ou encore des histogrammes.

**Exemple 1.2.6** *(Tableau de données symboliques - tiré de [3])*
Soient :

- $E = \{a_1, a_2, a_3, a_4\}$ un ensemble de quatre villes $(N = 4)$.
- $Y_1$ = le nombre d'étudiants (minimum et maximum des années 1990 – 1999, en milliers) d'où
  $\mathcal{B}_1 = \mathcal{J} = \{[\alpha, \beta] =$[min,max] tel que $0 \leq \alpha \leq \beta < \infty\}$.
  $Y_1$ est une variable de type intervalle.
- $Y_2$ = le spectre des opérateurs de téléphonie mobile dans une ville, c'est à dire un sous-ensemble de $\mathcal{Y}_2 = \{$B,M,P$\}$ comprenant les trois opérateurs B=Base, M=Mobistar et P=Proximus, ainsi que les pourcentages

de personnes couvertes par chacun des trois opérateurs. Donc, $\mathcal{B}_2$ est l'ensemble des distributions de fréquences sur $\mathcal{Y}_2$.
$Y_2$ est une variable modale (histogramme).

- $Y_3$ = la liste des salles de cinéma situées dans une ville, c'est un sous-ensemble de toutes les salles de cinéma.
$\mathcal{Y}_3$ = {Caméo, Carollyxood, Eldorado, Forum, Kinépolis; Le Parc, UGC,...}.
$\mathcal{B}_3 = \mathcal{P}(\mathcal{Y}_3)$. $Y_3$ est une variable multivaluée catégorique.

Nous obtenons le tableau de données symboliques suivant :

| | $Y_1$ | $Y_2$ | $Y_3$ |
|---|---|---|---|
| $a_1$ | [50, 57] | (B 0.3; M 0.3; P 0.4) | {Kinépolis, UGC} |
| $a_2$ | [6, 7] | (B 0.1; M 0.5; P 0.4) | {Le Parc, Paradiso} |
| $a_3$ | [20, 24] | (B 0.4; M 0.1; P 0.5) | {Kinépolis, Palace} |
| $a_4$ | [8, 9] | (B 0.2; M 0.3; P 0.4) | {Caméo, Eldorado, Forum} |

La deuxième ligne de ce tableau de données correspond à la description de la ville $a_2$.

### 1.2.7 Exemple de passage de données classiques à des données symboliques ( - tiré de [3])

Soit $\Omega$ un ensemble de $k$ individus décrits par $p$ variables classiques $\widetilde{Y}_1, \ldots, \widetilde{Y}_p$ d'espaces d'observation $\mathcal{Y}_1, \ldots, \mathcal{Y}_p$.

Les classes $C_i$ sont décrites par des variables symboliques $Y_1, \ldots, Y_p$ de telle façon que $Y_j(C_i)$ caractérise l'ensemble $\{\{\widetilde{Y}_j(k) \mid k \in C_i\} \subseteq \mathcal{Y}_j\}$ des valeurs observées pour $\widetilde{Y}_j$ à l'intérieur de la classe $C_i$.

Prenons une matrice de données classiques $\widetilde{X} = (x_{kj})$ de quinze individus et trois variables univaluées classiques :

- $\widetilde{Y}_1$ = marque de moto (Aprilia, BMW, Ducati, Harley Davidson, Honda, Kawasaki, Suzuki, Yamaha) ;
- $\widetilde{Y}_2$ = vitesse maximale (en kilomètres/heure) et
- $\widetilde{Y}_3$ = pays de production (Japon, Allemagne, Italie, Etats-Unis).

| Numéro de série $k$ | Marque $\widetilde{Y}_1$ | Vitesse Maximale $\widetilde{Y}_2$ | Fabricant $\widetilde{Y}_3$ | |
|---|---|---|---|---|
| 1 | Harley Davidson | 212 | Etats-Unis | $x'_1$ |
| 2 | Harley Davidson | 184 | Etats-Unis | $x'_2$ |
| 3 | Harley Davidson | 176 | Etats-Unis | $x'_3$ |
| 4 | Yamaha | 254 | Japon | $x'_4$ |
| 5 | Yamaha | 227 | Japon | $x'_5$ |
| 6 | Suzuki | 272 | Japon | $x'_6$ |
| 7 | Kawasaki | 263 | Japon | $x'_7$ |
| 8 | Honda | 312 | Japon | $x'_8$ |
| 9 | Honda | 289 | Japon | $x'_9$ |
| 10 | BMW | 259 | Allemagne | $x'_{10}$ |
| 11 | BMW | 284 | Allemagne | $x'_{11}$ |
| 12 | Aprilia | 275 | Italie | $x'_{12}$ |
| 13 | Aprilia | 318 | Italie | $x'_{13}$ |
| 14 | Ducati | 279 | Italie | $x'_{14}$ |
| 15 | Ducati | 294 | Italie | $x'_{15}$ |

Considérons les groupes

$$C_1 = \{1, 2, 3\} \quad C_2 = \{4, 5, 6, 7, 8, 9\} \quad C_3 = \{10, 11\} \quad C_4 = \{12, 13, 14, 15\}$$

de motos respectivement américaines, japonaises, allemandes et italiennes.

Après avoir agrégé les individus en quatre classes $C_1, C_2, C_3, C_4$, ce nouvel ensemble de classes (ou objets du second ordre) peut être décrit par une matrice de données symboliques $\underline{X} = (\xi_{ij})$ :

| | Marque $Y_1$ | Vitesse Maximale $Y_2$ | Fabricant $Y_3$ |
|---|---|---|---|
| $C_1$ | {Harley Davidson} | [176, 212] | Etats-Unis |
| $C_2$ | {Yamaha, Suzuki, Kawasaki, Honda} | [227, 312] | Japon |
| $C_3$ | {BMW} | [259, 284] | Allemagne |
| $C_4$ | {Aprilia, Ducati} | [275, 318] | Italie |

### 1.2.8 Dissimilarités entre objets symboliques

#### Pour des variables symboliques de type intervalle

Considérons une matrice de données symboliques $\underline{X}$ composée de $n$ objets décrits par $p$ variables de type intervalle. Nous avons :

$$\underline{X} = \begin{pmatrix} \xi_{11} & \xi_{12} & \cdots & \xi_{1p} \\ \xi_{21} & \xi_{22} & \cdots & \xi_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{n1} & \xi_{n2} & \cdots & \xi_{np} \end{pmatrix}$$

où $\xi_{kj} = Y_j(k) = [\alpha_{kj}, \beta_{kj}]$ est la description de la $j^{ème}$ composante de l'objet $k \in E$.

Nous allons maintenant définir une mesure de dissimilarité sur $E$ à partir de $p$ indices de dissimilarité sur les $\mathcal{B}_j$. Ainsi :

$$\begin{array}{rccc} \delta_j : & \mathcal{B}_j \times \mathcal{B}_j & \to & \mathbb{R}^+ \\ & (\xi_{kj}, \xi_{lj}) & \rightsquigarrow & \delta_j(\xi_{kj}, \xi_{lj}) \end{array}$$

A partir de deux intervalles $\xi_{kj} = [\alpha_{kj}, \beta_{kj}]$ et $\xi_{lj} = [\alpha_{lj}, \beta_{lj}]$, nous pouvons définir trois types de distances.

- La distance de Haussdorff :

$$\delta_j(\xi_{kj}, \xi_{lj}) = \max\{|\alpha_{kj} - \alpha_{lj}|, |\beta_{kj} - \beta_{lj}|\}$$

  On prend le maximum entre la valeur absolue de la différence des bornes inférieures des deux intervalles et la valeur absolue de la différence de ses bornes supérieures.

- La distance $L_1$ :

$$\delta_j(\xi_{kj}, \xi_{lj}) = |\alpha_{kj} - \alpha_{lj}| + |\beta_{kj} - \beta_{lj}|$$

  C'est la somme des valeurs absolues des différences entre les bornes inférieures et les bornes supérieures.

- La distance $L_2$ :

$$\delta_j(\xi_{kj}, \xi_{lj}) = (\alpha_{kj} - \alpha_{lj})^2 + (\beta_{kj} - \beta_{lj})^2$$

  C'est la somme des carrés des différences entre les bornes inférieures et les deux bornes supérieures.

Afin de se ramener à une dissimilarité sur l'ensemble $E$ des individus, nous sommes amenés à définir :

$$\begin{array}{rccl} d: & E \times E & \to & \mathbb{R}^+ \\ & (k,l) & \rightsquigarrow & d(k,l) = \left(\sum_{j=1}^{p} \delta_j^2(\xi_{kj}, \xi_{lj})\right)^{\frac{1}{2}} \end{array}$$

où $\delta_j$ est une des distances définies précédemment.

### Pour des variables multivaluées

$E$ est un ensemble de $n$ objets décrits par $p$ variables multivaluées $Y_1, \ldots, Y_p$ dont les espaces d'observation sont $\mathcal{Y}_1, \ldots, \mathcal{Y}_p$.

Soit $m_j$ le nombre de catégories prises par $Y_j$.

La fréquence $q_{j,k}(c_s)$ associée à chaque catégorie $c_s (s = 1, \ldots, m_j)$ de $Y_j$ pour l'objet $k$ est donnée par

$$q_{j,k}(c_s) = \begin{cases} \frac{1}{|Y_j(k)|} & \text{si} \quad c_s \in Y_j(k) \\ 0 & \text{sinon.} \end{cases}$$

La représentation symbolique de l'objet $k \in E$ est

$$k = ((q_{1,k}(c_1), \ldots, q_{1,k}(c_{m_1})), \ldots, (q_{p,k}(c_1), \ldots, q_{p,k}(c_{m_p}))).$$

La matrice originale $\underline{X} = (Y_j(k))$ est transformée en une matrice de fréquence $\widetilde{X}$ :

| | $Y_1$ | | | ... | $Y_p$ | | |
|---|---|---|---|---|---|---|---|
| | 1 | ... | $m_1$ | ... | 1 | ... | $m_p$ |
| 1 | $q_{1,1}(c_1)$ | ... | $q_{1,1}(c_{m_1})$ | ... | $q_{p,1}(c_1)$ | ... | $q_{p,1}(c_{m_p})$ |
| ⋮ | ⋮ | | ⋮ | | ⋮ | | ⋮ |
| $k$ | $q_{1,k}(c_1)$ | ... | $q_{1,k}(c_{m_1})$ | ... | $q_{p,k}(c_1)$ | ... | $q_{p,k}(c_{m_p})$ |
| ⋮ | ⋮ | | ⋮ | | ⋮ | | ⋮ |
| $n$ | $q_{1,n}(c_1)$ | ... | $q_{1,n}(c_{m_1})$ | ... | $q_{p,n}(c_1)$ | ... | $q_{p,n}(c_{m_p})$ |

où $\forall k \in E$, et $\forall j \in \{1, \ldots, p\}$, $\sum_{i=1}^{m_j} q_{j,k}(c_i) = 1$.

**Exemple 1.2.7**
Considérons la matrice suivante

| $k$ | $Y$ : couleur |
|---|---|
| 1 | {rouge} |
| 2 | {bleu, vert, rouge} |
| 3 | {bleu jaune} |

où $Y(k)$ : "la couleur de l'objet $k$ " est une variable multivaluée catégorique dont l'espace d'observation est $\mathcal{Y} = \{\text{bleu, jaune, rouge, vert}\}$.

Nous codons alors cette matrice de la façon suivante :

| $k$ | couleur | | | |
|---|---|---|---|---|
| | bleu | jaune | rouge | vert |
| 1 | 0 | 0 | 1 | 0 |
| 2 | 1/3 | 0 | 1/3 | 1/3 |
| 3 | 1/2 | 1/2 | 0 | 0 |

□

Soit $\delta_j$ une distance sur $\mathcal{B}_j$ :

$$\begin{array}{rccl} \delta_j : & \mathcal{B}_j \times \mathcal{B}_j & \to & \mathbb{R}^+ \\ & (x_{kj}, x_{lj}) & \rightsquigarrow & \delta_j(x_{kj}, x_{lj}). \end{array}$$

Les distances $L_1$ et $L_2$ sur $\mathcal{B}_j$ sont définies par :

$$\delta_j(\xi_{kj}, \xi_{lj}) = \sum_{i=1}^{|\mathcal{Y}_j|} |q_{j,k}(c_i) - q_{j,l}(c_i)| \quad \text{et} \quad \delta_j(\xi_{kj}, \xi_{lj}) = \sum_{i=1}^{|\mathcal{Y}_j|} (q_{j,k}(c_i) - q_{j,l}(c_i))^2$$

La distance de *de Carvalho* est définie par :

$$\delta_j(\xi_{kj}, \xi_{lj}) = \sum_{i=1}^{|\mathcal{Y}_j|} (\gamma q_{j,k}(c_i) + \gamma' q_{j,l}(c_i))$$

où

- $\gamma = \begin{cases} 1 & \text{si} \quad c_i \in Y_j(k) \text{ et } c_i \notin Y_j(x_l) \\ 0 & \text{sinon} \end{cases}$
- $\gamma' = \begin{cases} 1 & \text{si} \quad c_i \notin Y_j(k) \text{ et } c_i \in Y_j(x_l) \\ 0 & \text{sinon} \end{cases}$

On combine les $p$ indices de dissimilarité $\delta_1, \ldots, \delta_p$ définis sur les domaines $\mathcal{B}_j$ en une dissimilarité globale sur $E$

$$\begin{array}{rccl} d: & E \times E & \to & \mathbb{R}^+ \\ & (k,l) & \rightsquigarrow & d(k,l) = \left(\sum_{j=1}^{p} \delta_j^2(x_{kj}, x_{lj})\right)^{1/2} \end{array}$$

où $\delta_j$ est une des mesures de dissimilarité définies ci-dessus.

**Remarque :**
Le cas des variables modales est similaire au cas des variables multivaluées. Les fréquences $q_{j,k}(c_s)$ sont simplement remplacées par les valeurs de la distribution $\pi_{j,k}$ associées à chacune des catégories de $Y_j(k)$.

# Chapitre 2

# Indices, compatibilités et hiérarchies

## 2.1 Définitions

**Définition 2.1.1**

Un **indice de dissimilarité** $d$ sur $\Omega$ est une application

$$\begin{array}{rcl} d : \Omega \times \Omega & \rightarrow & \mathbb{R}^+ \\ (w, w') & \mapsto & d(w, w') \end{array}$$

qui satisfait aux deux propriétés suivantes :

1. $d$ est symétrique : $\forall (w, w') \in \Omega \times \Omega,\ d(w, w') = d(w', w)$
2. $d(w, w') = 0$ si $w \equiv w'$.

**Définition 2.1.2**

Un **indice de distance** $d$ sur $\Omega$ est un indice de dissimilarité telle que

$$\forall (w, w') \in \Omega \times \Omega,\ d(w, w') = 0 \Rightarrow w \equiv w' \ .$$

**Définition 2.1.3**

Une **distance** $d$ sur $\Omega$ est un indice de distance qui vérifie en plus l'inégalité triangulaire :

$$\forall w, w', w'' \in \Omega,\ d(w, w') \leq d(w, w'') + d(w', w'') \ .$$

**Définition 2.1.4**

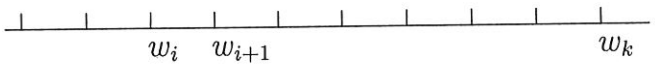
Une **distance ultramétrique** sur $\Omega$ est une distance qui vérifie en plus l'inégalité ultramétrique :

$$\forall w, w', w'' \in \Omega,\ d(w, w') \leq \max\left(d(w, w''), d(w', w'')\right) .$$

## 2.2 Les compatibilités entre ordres et indices de dissimilarité

**Définition 2.2.1**

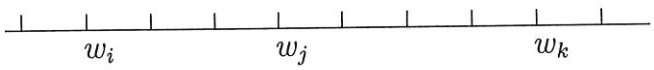
Un indice de dissimilarité $d$ et un ordre $\theta$ sont **faiblement compatibles** si et seulement si pour tout triplet ordonné selon $\theta$ et ayant deux éléments consécutifs, la distance des éléments consécutifs est inférieure à la distance des éléments extrêmes.




$$d(w_i, w_{i+1}) \leq d(w_i, w_k)$$

**Définition 2.2.2**

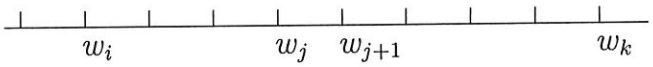
$\theta$ et $d$ sont **compatibles** si et seulement si pour tout triplet ordonné selon $\theta$, la distance entre les éléments extrêmes est supérieure à la distance entre une extrémité quelconque et l'élément intermédiaire.




$$d(w_i, w_k) \geq \max(d(w_i, w_j), d(w_j, w_k))$$

**Définition 2.2.3**

$\theta$ et $d$ sont **semi-compatibles** si et seulement si pour tout quadruplet ordonné selon $\theta$ et dont les éléments intermédiaires sont consécutifs, la distance des éléments extrêmes est supérieure à la distance des éléments consécutifs.




$$d(w_i, w_k) \geq d(w_j, w_{j+1})$$

## 2.3 Les matrices de Robinson, Sur-Diagonales Rectangles et Sur-Diagonales Dominées

Pour le terme "sur-diagonale rectangle", nous utiliserons l'abbréviation SDR et pour "sur-diagonale dominée", l'abbréviation SDD.

Soit $D$ la matrice de dissimilarité associée à l'indice $d$ :

$$D = \{d_{ij}\} \qquad i, j = 1, \ldots, n$$

Cette matrice étant symétrique, on pourra établir les définitions de matrice de Robinson, SDR et SDD en ne considérant que la partie triangulaire supérieure de $D$.

**Définition 2.3.1**

Une matrice est dite de **Robinson** si et seulement si les termes des lignes et des colonnes sont croissants à partir de chaque terme de la diagonale.

Un exemple est donné par la matrice suivante :

$$\begin{pmatrix} 0 & 2 & 4 & 6 \\ 2 & 0 & 4 & 5 \\ 4 & 4 & 0 & 1 \\ 6 & 5 & 1 & 0 \end{pmatrix}$$

Considérons la matrice triangulaire supérieure déduite de $D$ en excluant la diagonale de $D$. La sur-diagonale est alors la plus grande diagonale de cette matrice. Par exemple, dans la matrice de Robinson ci-dessus, la sur-diagonale est : 2 4 1.

A chaque terme de la sur-diagonale, on peut associer un rectangle dont les côtés sont formés de la ligne et de la colonne contenues dans la matrice triangulaire supérieure et issues de ce terme.

Pour le même exemple de la matrice de Robinson, le rectangle (ici un carré) issu du terme 4 de la sur-diagonale est $\begin{matrix} 4 & 6 \\ 4 & 5 \end{matrix}$.

**Définition 2.3.2**

Une matrice est dite **SDR**, c'est-à-dire sur-diagonale "rectangle", si chaque terme de la sur-diagonale est inférieur aux termes du rectangle qui lui est associé.

Un exemple est donné par la matrice suivante :

$$\begin{pmatrix} 0 & 2 & 6 & 5 \\ 2 & 0 & 4 & 7 \\ 6 & 4 & 0 & 1 \\ 5 & 7 & 1 & 0 \end{pmatrix}$$

**Définition 2.3.3**

Une matrice est dite **SDD**, c'est-à-dire sur-diagonale "dominée", si dans la matrice triangulaire supérieure associée à $D$, les termes des lignes et des colonnes sont plus grands que le terme de la sur-diagonale qu'elles contiennent.

Un exemple est donné par la matrice suivante :

$$\begin{pmatrix} 0 & 2 & 5 & 3 \\ 2 & 0 & 4 & 6 \\ 5 & 4 & 0 & 1 \\ 3 & 6 & 1 & 0 \end{pmatrix}$$

On peut résumer ces trois définitions de manière plus formelle par le tableau 2.1 , (où les intervalles de variation de $i$, $j$ et $l$ entre 1 et $n$ se déduisent immédiatemment des différentes formules).

**Pour $i \leq j$ :**

$$\text{Robinson} \iff \begin{cases} \text{Condition lignes : } d_{ij} \leq d_{ij+1} \\ \text{Condition colonnes : } d_{ij} \leq d_{i-1j} \end{cases}$$

$$\text{SDR} \iff d_{jj+1} \leq d_{il} \quad j+1 \leq l$$

$$\text{SDD} \iff \begin{cases} \text{Condition lignes : } d_{ii+1} \leq d_{ij+1} \\ \text{Condition colonnes : } d_{j-1j} \leq d_{i-1j} \end{cases}$$

TAB. 2.1 – Les matrices Robinson, SDR et SDD

Il résulte facilement de ces trois définitions que l'ensemble des matrices de Robinson est inclus dans l'ensemble des matrices SDR qui est lui-même inclus dans l'ensemble des matrices SDD.

Notons $M(d, \theta)$, la matrice de dissimilarité associée à $d$ et dont les lignes et les colonnes sont rangées selon l'ordre $\theta$. Alors on a, en plus, les trois propriétés suivantes :

| | | |
|---|---|---|
| $M(d,\theta)$ est Robinson | $\iff$ | $d$ et $\theta$ sont compatibles |
| $M(d,\theta)$ est SDR | $\iff$ | $d$ et $\theta$ sont semi-compatibles |
| $M(d,\theta)$ est SDD | $\iff$ | $d$ et $\theta$ sont faiblement compatibles |

## 2.4 Hiérarchies, ultramétriques et ordres

**Définition 2.4.1**

Soit $\Omega$ un ensemble fini, $H$ un ensemble de parties (appelées paliers) de $\Omega$, $H$ est une **hiérarchie** sur $\Omega$ si et seulement si :

1. $\Omega \in H$ et $\emptyset \notin H$
2. $\forall w \in \Omega,\ \{w\} \in H$
3. $\forall (h, h') \in H^2$ on a $h \cap h' = \emptyset \Rightarrow h \subset h'$ ou $h' \subset h$.

**Définition 2.4.2** *Première définition de croisement.*

Un ordre donne lieu à un **croisement** pour la hiérarchie $H$ si et seulement si il existe $h \in H$ et $(w_i, w_j, w_l) \in \Omega$, avec $w_i < w_j < w_l$ selon $\theta$, tels que $w_i, w_l \in h$ et $w_j \notin h$.

**Proposition 2.4.1** *Il existe une bijection entre l'ensemble des hiérarchies indicées et l'ensemble des ultramétriques.*

Voir référence [2].

Notons $\delta_H$ l'ultramétrique associée à une hiérarchie indicée notée $H$ par la bijection.

**Proposition 2.4.2** *Une condition nécessaire et suffisante pour qu'un ordre $\theta$ ne donne pas lieu à un croisement pour une hiérarchie indicée $H$ est que $\delta_H$ et $\theta$ soient faiblement compatibles.*

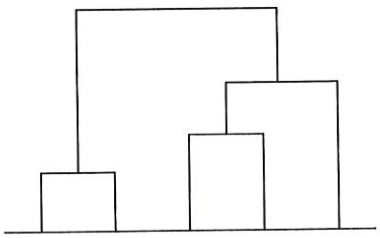
Voir référence [2].

**Proposition 2.4.3** *Si $\delta$ est une ultramétrique, il existe toujours un ordre $\theta$ (non unique) tel que $\delta$ et $\theta$ sont faiblement compatibles.*

**Preuve :** Il est toujours possible de définir un ordre $\theta$ sans croisement à partir d'une hiérarchie $H$ en associant, à chaque palier, un ordre sur les plus grands paliers dont il est la réunion. Il faut procéder ainsi à partir du palier identique à $\Omega$ en allant jusqu'aux paliers ne contenant qu'un singleton. De cette manière, on obtient un ordre sur les individus qui ne peut comporter de croisement.

Soit $H$, la hiérarchie indicée associée à $\delta$ qui existe grâce à la proposition 2.4.1. Pour cette hiérarchie, il existe donc un ordre $\theta$ ne donnant pas lieu à un croisement. On a donc, par la proposition 2.4.2, que $\delta$ et $\theta$ sont faiblement compatibles.

Voici un exemple de hiérarchie telle que le $\delta$ et le $\theta$, associés à cette hiérarchie, sont faiblement compatibles.



□

**Proposition 2.4.4** *Si $\delta$ est une ultramétrique et si $M(\delta, \theta)$ est SDD alors $M(\delta, \theta)$ est une matrice de Robinson.*

Voir référence [2].

**Preuve :** On va montrer que $M(\delta, \theta)$ est Robinson en vérifiant les conditions "lignes" et "colonnes" (données dans le tableau 2.1) qui caractérisent une matrice de Robinson.

Comme $M(\delta, \theta)$ est SDD, on a que $\delta_{ij+1} \geq \delta_{jj+1}$ pour $i \leq j$ . De plus, $\delta$ est une ultramétrique donc on a $\delta_{ij} \leq \max(\delta_{ij+1}, \delta_{jj+1})$ pour $i \leq j$ , d'où on satisfait la condition "lignes" :

$$\delta_{ij} \leq \delta_{ij+1} \quad \text{pour} \quad i \leq j \ .$$

De même, $M(\delta, \theta)$ SDD implique qu'on a $\delta_{i-1j} \geq \delta_{j-1j}$ pour $i \leq j$ et en utilisant le fait que $\delta$ est une ultramétrique, on a aussi $\delta_{ij} \leq \max(\delta_{i-1j}, \delta_{i-1i})$ pour $i \leq j$ . Dès lors, la condition "colonnes" est également vérifiée car on a bien que

$$\delta_{ij} \leq \delta_{i-1j} \quad \text{pour} \quad i \leq j \ .$$

Les deux conditions pour que $M(\delta, \theta)$ soit Robinson sont satisfaites.

**Remarque :** Ce résultat est assez important dans le sens que si $\delta$ est une ultramétrique, alors les différents types de compatibilité sont équivalents.

**Définition 2.4.3**

Si $\delta$ est une ultramétrique et $\theta$ l'ordre tel que $\delta$ et $\theta$ soient faiblement compatibles, la matrice $M(\delta, \theta)$ est alors dite **ultramétrique**. Ses éléments satisfont l'inégalité ultramétrique :

$$\forall(i, j, k) \ : \ \delta(w_i, w_k) \ \leq \ \max\left(\delta(w_i, w_j), \delta(w_j, w_k)\right) \ .$$

**Conséquence :** De plus, elle est Robinson par la proposition 2.4.4 car $M(\delta, \theta)$ est SDD.

## 2.5 Représentation visuelle

Il faut faire remarquer que la représentation hiérarchique donne en fait une image visuelle du contenu d'une matrice ultramétrique, elle se base sur la compatibilité entre un ordre et une ultramétrique.

Puisque la représentation d'une matrice ultramétrique est possible, il est naturel de penser à représenter visuellement les autres types de compatibilité et plus précisement les matrices Robinson, SDR et SDD.

Comme le montre le tableau suivant, ces différentes compatibilités sont une extension de la notion de compatibilité entre un ordre et une ultramétrique. Leurs représentations visuelles étendent également la représentation hiérarchique.

C'est ainsi que l'on aboutit à une nouvelle forme de représentation visuelle qui a l'allure d'une *pyramide*. Remarquons que plus la matrice de dissimilarité représentée se rapproche d'une matrice ultramétrique, plus la forme de la pyramide se rapproche d'une hiérarchie.

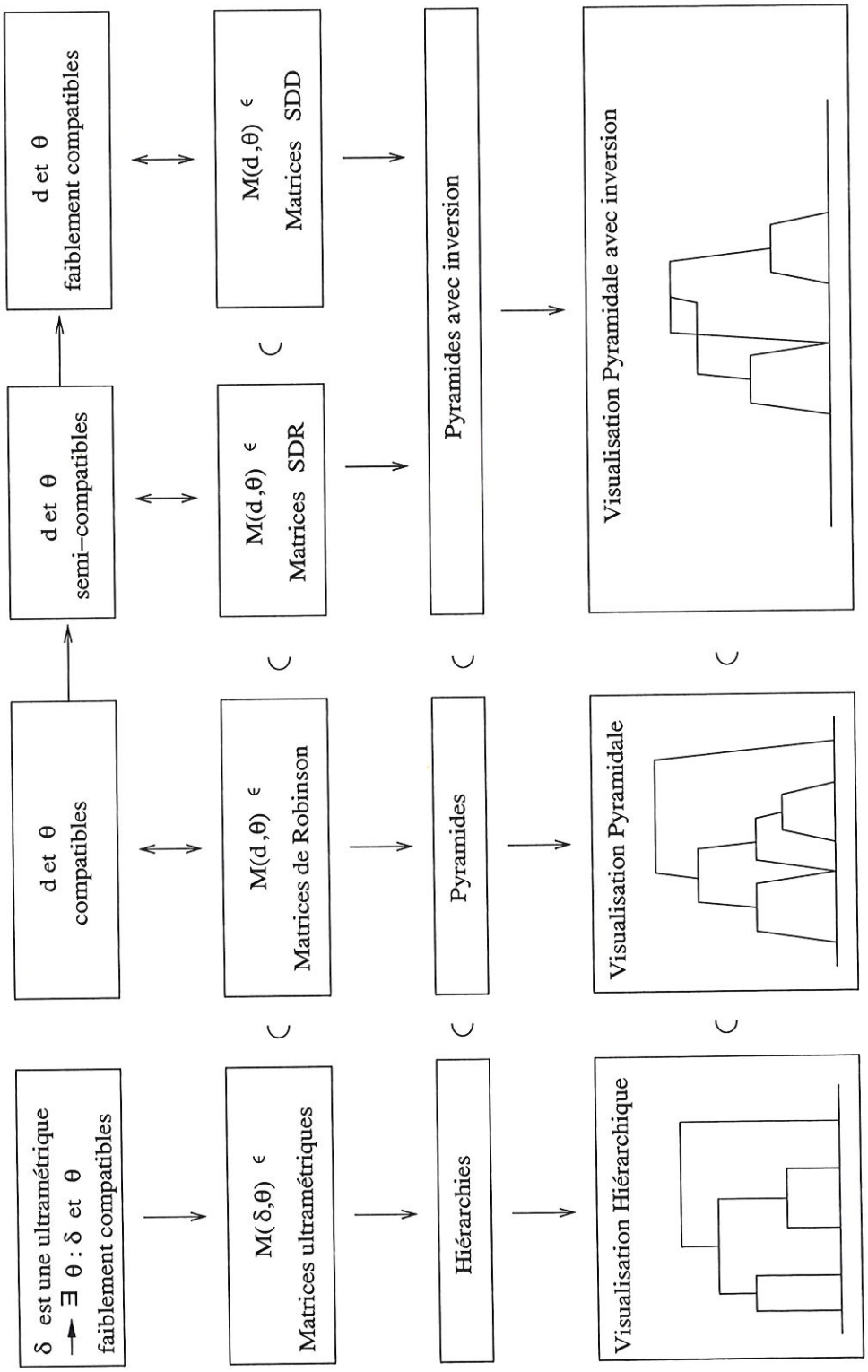



FIG. 2.1 – Illustration des compatibilités

# Chapitre 3

# Les Pyramides

## 3.1 Définition et proposition

La section précédente énoncait déjà une certaine approche des pyramides en comparaison avec les hiérarchies. Ici, nous allons nous intéresser à la définition exacte des pyramides.

Avant de donner cette définition, il reste à introduire une notion très utile qui est celle de *compatibilité* entre un ordre $\theta$ sur $\Omega$ et un ensemble $P$ de parties de $\Omega$.

**Définition 3.1.1**

Une partie $h$ de $P$ est **connexe selon** $\theta$ si pour $w'$ et $w''$, les bornes (le plus petit et le plus grand élément) de $h$ selon $\theta$, on a la condition suivante :

$$\{w \text{ compris entre } w' \text{ et } w'' \text{ selon } \theta\} \Longleftrightarrow \{w \in h\} \ .$$

**Définition 3.1.2**

Un ordre $\theta$ est **compatible** avec un ensemble $P$ de parties de $\Omega$ si tout élément $h \in P$ est connexe selon $\theta$.

**Définition 3.1.3**

Soit $\Omega$ un ensemble fini, $P$ un ensemble de parties non vides (appelées paliers) sur $\Omega$, $P$ est une **pyramide** sur $\Omega$ si et seulement :

1. $\Omega \in P$ ; le plus grand palier contient tous les individus,
2. $\forall w \in \Omega$, $\{w\} \in P$ ; les plus petits paliers sont les singletons,
3. $\forall (h, h') \in P^2$ on a $h \cap h' = \emptyset$ ou $h \cap h' \in P$,
4. Il existe un ordre $\theta$ compatible avec $P$.

**Exemple 3.1.1** *(Pyramide)*
Soit

$$\Omega = \{w_1, w_2, w_3\} \text{ et}$$
$$P_1 = \{\{w_1\}, \{w_2\}, \{w_3\}, \{w_1, w_2\}, \{w_2, w_3\}, \Omega\} ,$$
$$P_2 = P_1 \cup \{w_1, w_3\} .$$

Prenons $\theta$ comme étant l'ordre $w_1, w_2, w_3$.

On remarque immédiatemment que $P_1$ est une pyramide tandis que $P_2$ n'est est pas une car la quatrième condition n'est pas vérifiée. En effet la classe $\{w_1, w_3\}$ ne contient pas $w_2$ alors qu'il est compris entre $w_1$ et $w_3$ selon $\theta$.

**Définition 3.1.4** *Deuxième définition de croisement.*

Un ordre $\theta$ donne lieu à un **croisement** pour une pyramide $P$ s'il existe un palier de $P$ qui n'est pas connexe selon $\theta$, c'est-à-dire si $\theta$ n'est pas compatible avec $P$.

**Proposition 3.1.1** *L'ensemble des hiérarchies est inclus dans l'ensemble des pyramides.*

**Preuve :** Montrons que les hiérarchies satisfont aux quatres conditions de la définition de pyramide.
Soit $H$, une hiérarchie :

- les 2 premières conditions à satisfaire sont identiques à celles de la définition d'une hiérarchie.

- Pour vérifier la troisième condition, rappelons-nous que pour que $H$ soit une hiérarchie il faut que

$$\forall\ (h, h') \in H \times H,\ h \cap h' = \emptyset \text{ ou } h \cap h' = h \text{ ou } h \cap h' = h',$$

or $\emptyset$, $h$ et $h' \in H$, donc la troisième condition est vérifiée.

- Afin de prouver la quatrième condition, on va construire un ordre sur $\Omega$ induit par $H$ et montrer qu'il est compatible avec $H$.
Pour construire cet ordre, on part du fait que les paliers de $H$ sont soit emboîtés, soit d'intersection vide. On prend $\Omega$, on choisit un ordre sur les plus grandes parties de $H$ contenues dans $\Omega$.
On recommence le procédé avec chacune de ces parties en choisissant un ordre sur les plus grandes parties qu'elles contiennent et ainsi de suite jusqu'aux singletons qui respectent l'ordre induit par ce procédé et qu'on note $\theta$.
Montrons que cet ordre est compatible avec $H$. En effet, soit $h \in H$ et soit $(w', w'')$ les extrêmités de $h$ selon $\theta$, tous les éléments compris entre $w'$ et $w''$ appartiennent par construction même à des parties $h_l \in H$ incluses dans $h$ et n'appartiennent qu'à celles-ci donc $h$ est connexe selon $\theta$. La compatibilité entre $\theta$ et $H$ est donc vérifiée.

□

## 3.2 Visualisation d'une pyramide

### 3.2.1 Notions de successeurs, prédécesseurs et niveaux

Soit $P$ une pyramide et $\Omega$ l'ensemble des individus :

**Définition 3.2.1**

On dira que $h \in P$ est **successeur** de $h' \in P$ si $h \subset h'$ au sens strict et (sauf si $h'$ est un singleton ou si $h = \Omega$) s'il n'existe pas $h'' \neq h$ et $h'$ tel que $h \subset h'' \subset h'$ au sens strict.
On peut dire aussi que $h'$ est **prédécesseur** de $h$.

**Définition 3.2.2**

Sachant que $\Omega \in P$, l'ensemble des successeurs de $\Omega$ forme un recouvrement de $\Omega$ puisque $P$ contient les singletons, un tel recouvrement est appelé **niveau de la pyramide**. L'ensemble des successeurs des paliers qui forment ce recouvrement forme un nouveau **niveau** qui est également un recouvrement. On peut passer de cette manière d'un niveau au suivant jusqu'à atteindre un niveau composé uniquement de singletons car d'un niveau à l'autre, la taille des paliers va en se réduisant.

Pour permettre la visualisation d'une pyramide, il faut les quatres aspects suivants :

**(a)** Préciser comment les paliers s'imbriquent les uns dans les autres.

**(b)** Construire un ordre sur les singletons qui soit compatible avec la pyramide.

**(c)** Dire comment se fait la représentation graphique.

**(d)** Indicer la pyramide, c'est-à-dire associer une hauteur à chaque palier, (mais cela ne se fera qu'au chapitre suivant).

### 3.2.2 Nombre maximum de prédécesseurs d'un palier d'une pyramide

Il faut introduire encore quelques notions avant de présenter les différentes propositions concernant les paliers d'une pyramide.

**Définition 3.2.3**

• Deux paliers $h$ et $h'$ sont dits **connexes** s'il existe une suite de paliers $h_1, \ldots, h_q$ telle que $h_1 = h$, $h_q = h'$ et $h_i \cap h_{i+1} \neq \emptyset$ pour $i = 1, \ldots, q-1$.

• Une **partie connexe** est un ensemble de paliers connexes entre eux.

• L'ensemble des plus grandes parties connexes associées à chaque niveau, appelées **classes connexes**, forme une partition de $\Omega$ (parfois réduite à un seul élément si tous les paliers du niveau sont connexes entre eux).

• La relation $R$ : $hRh' \iff \{h$ et $h'$ sont connexes$\}$ est réflexive, symétrique et transitive.

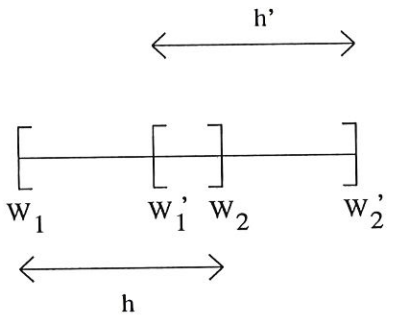
**Définition 3.2.4**

Soit $\theta$ un ordre compatible avec $P$. Un palier $h$ est dit **à gauche** (respectivement **à droite**) d'un palier $h'$ si parmi les paliers qui contiennent le plus petit (respectivement le plus grand) élément de $h'$ selon $\theta$, $h$ a une intersection de plus grande taille avec $h'$.

**Proposition 3.2.1** *Soient $h$ et $h'$ deux paliers d'une même classe connexe, si $h$ est à gauche (respectivement à droite) de $h'$ alors $h'$ est l'unique élément à droite (respectivement à gauche) de $h$.*

**Preuve :** Sans perte de généralité, on peut dire que tous les paliers de $P$ sont connexes selon un ordre $\theta$ compatible avec $P$.

Soient $w_1$ et $w_2$ (respectivement $w'_1$ et $w'_2$) les éléments à gauche et à droite de $h$ ( respectivement de $h'$) selon $\theta$. Si $h$ est à gauche de $h'$, la seule configuration possible est celle de la figure suivante :




En effet, $w_2$ ne peut être à droite de $w'_2$ car alors $h'$ serait inclus dans $h$ et ne pourrait être successeur du niveau précédent, c'est-à-dire $h$ et $h'$ n'appartiendraient pas au même niveau. De plus, $w_2$ ne peut pas être à gauche de $w'_1$ car sinon l'intersection serait vide et cela contredirait le fait que $h$ est à gauche de $h'$. Enfin, $w_1$ ne peut pas être compris entre $w'_1$ et $w_2$ car alors $h$ serait inclus dans $h'$ et ne pourrait être à son tour successeur du niveau précédent.

Sachant que $h$ est à gauche de $h'$, on va maintenant pouvoir montrer facilement que $h'$ est à droite de $h$. Pour cela, supposons qu'il existe un palier

$h''$ contenant $w_2$ (pour contenir le plus grand élément de $h$) et d'intersection plus grande que $h' \cap h$ avec $h$.
Si son plus petit élément $w_1''$ est inférieur à $w_1'$, alors son plus grand élément $w_2''$ doit être inférieur ou égal à $w_2$ (donc égal). En effet, si $w_2''$ était plus grand que $w_2$, l'intersection $h' \cap h''$ serait plus grande que $h' \cap h$ alors que par hypothèse, $h$ est à gauche de $h'$ et donc de plus grande intersection avec $h'$. Donc comme $w_2''$ est égal à $w_2$, $h''$ est inclus dans $h$ et ne peut donc être successeur du niveau précédent.

L'élément $w_1''$ ne peut donc être inférieur à $w_1'$ et est donc supérieur ou égal à cet élément. Le palier $h'$ est donc bien une partie de plus grande intersection avec $h$.

Il nous reste à montrer que $h'$ est l'unique palier de plus grande intersection avec $h$.

Supposons qu'il en existe un autre que l'on note $h''$, son plus petit élément doit alors être égal à $w_1'$ et pour le plus grand élément on a soit $w_2'' < w_2'$, soit $w_2'' > w_2'$, afin que $h''$ soit différent de $h'$.

Dans le premier cas, $h'' \subset h'$ strictement et $h''$ ne peut donc être un successeur du niveau précédent. Dans le second cas, c'est $h' \subset h''$ et donc $h'$ qui ne peut être successeur du niveau précédent.

□

**Proposition 3.2.2** *Chaque palier d'une pyramide a au maximum deux prédécesseurs.*

**Preuve :** Tout palier $h$ de $P$ a au moins un prédécesseur. Soit un palier $h$, s'il est dans plusieurs paliers prédécesseurs alors il est identique à leur plus grande intersection sinon il serait inclus dans cette intersection et ne pourrait être successeur du niveau précédent. En effet, la plus grande intersection est un successeur puisqu'elle appartient à $P$ par définition d'une pyramide et elle n'appartient pas au niveau précédent puisqu'elle est strictement incluse dans les paliers prédécesseurs qui la créent.

Finalement, la plus grande intersection est obtenue par deux paliers uniques car d'après la proposition précédente, sachant que les paliers sont à gauche et à droite l'un de l'autre, ils sont uniques. Donc $h$ a au maximum deux prédécesseurs.

□

### 3.2.3 Construction d'un ordre compatible avec une pyramide

**Proposition 3.2.3** *Une condition nécessaire et suffisante pour qu'un ordre $\theta$ sur $\Omega$ soit compatible avec une pyramide $P$ est que toute classe connexe $C$, ordonnée selon $\theta$, soit ordonnée selon une suite unique $h_1, \ldots, h_q$ caractérisée par les deux propriétés suivantes :*
$\forall i = 2, \ldots, q-1$ :

*(a) $h_{i-1}$ et $h_{i+1}$ sont parmi les paliers de $C$ , ceux qui sont de plus grande intersection avec $h_i$, (i.e. les paliers à gauche et à droite de $h_i$).*

*(b) $h_i \cap h_{i+1}$ est différent de $h_i$ et de $h_{i+1}$.*

**Preuve :** Voir référence [1]

**Remarque :** Il existe un algorithme constructif d'un ordre compatible avec une pyramide, ceci signifie qu'il est toujours possible d'en contruire un et dès lors qu'il en existe toujours au moins un. Nous ne développons pas cet algorithme ni la proposition 3.2.3 car cela dépasserait l'objectif de ce travail.

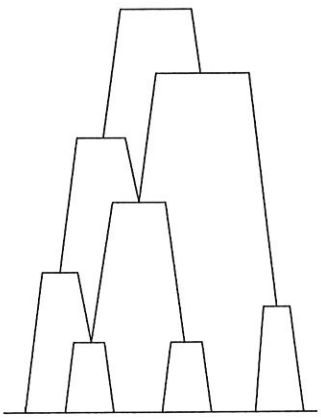
### 3.2.4 Représentation graphique

En ce qui concerne la représentation graphique, on va utiliser les trois caractéristiques des pyramides suivantes :

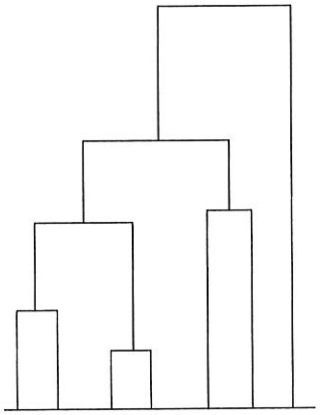
1. deux paliers peuvent être d'intersection non vide.
2. il existe au moins un ordre pour lequel chaque palier est connexe.
3. chaque palier a au maximum deux prédécesseurs.

Chaque palier est représenté par un segment horizontal, il est relié à ses prédécesseurs par des lignes obliques. Chaque ligne oblique relie le milieu du segment associé à un palier à l'extrêmité du segment associé à son ou (ses) prédécesseur(s).

Les deux figures suivantes illustrent bien les différences de représentation entre une pyramide et une hiérarchie.



PYRAMIDE



HIERARCHIE

# Chapitre 4

# Les Pyramides Indicées et les Indices Pyramidaux

## 4.1 Indiçage d'une pyramide

### 4.1.1 Pyramides indicées

**Définition 4.1.1**

Une pyramide **indicée** est un couple $(P, f)$ où $P$ est une pyramide et $f$ une application de $P$ dans $\mathbb{R}^+$ telle que :

1. $f(h) = 0 \Leftrightarrow h$ ne contient qu'un seul élément.
2. $\forall (h, h') \in P \times P$, $h \subset h'$ (inclusion stricte) $\Rightarrow f(h) \leq f(h')$.

**Définition 4.1.2**

- Une pyramide indicée est **indicée au sens large** si

$$\left.\begin{array}{l} h \subset h' \text{ (strictement)} \\ f(h) = f(h') \end{array}\right\} \Longrightarrow \exists\, h_1 \text{ et } h_2 \in P \text{ et } \neq h : h = h_1 \cap h_2 \ .$$

(Si $h$ pouvait être égal à $h_1$ ou $h_2$, la condition serait toujours vérifiée avec $h = h' \cap h$).

- Une pyramide est **indicée au sens strict** si

$$h \subset h' \text{ (strictement)} \Rightarrow f(h) < f(h') \ .$$

- La quantité $f(h)$ est appelée **hauteur** du palier $h$.

**Exemple 4.1.1** *(Différentes pyramides indicées)*

- Soit :

$\Omega = \{a, b, c, d\}$

$P_1 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{abc\}, \{bc\}, \{bcd\}, \Omega\}$

$f(P_1) = \{0, 0, 0, 0, 2, 1, 2, 3\}$

$P_1$ est une pyramide indicée au sens strict. (Voir figure $(a)$).

- Soit :

$P_2 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{abc\}, \{bc\}, \{bcd\}, \Omega\}$

$f(P_2) = \{0, 0, 0, 0, 1, 1, 2, 3\}$

$P_2$ est une pyramide indicée au sens large puisque les seuls paliers pour lesquels $h \subset h'$ et $f(h) = f(h')$ sont les suivants :

$$h_1 = \{bc\} \subset h_2 = \{abc\} \quad et \quad f(h_1) = f(h_2)$$

$$\Longrightarrow \exists\ \{abc\}, \{bcd\} : \{abc\} \cap \{bcd\} = \{bc\} \ ,$$

la condition est donc bien satisfaite. (Voir figure $(b)$).

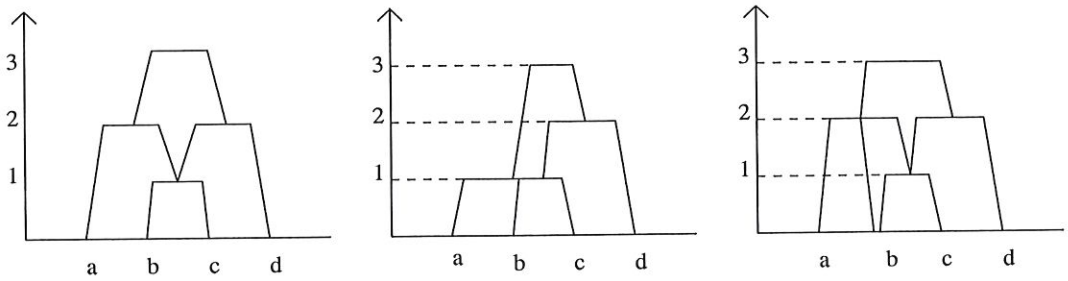
- Soit :

$P_3 = P_1 \cup \{ab\}$

$P_3 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{ab\}, \{abc\}, \{bc\}, \{bcd\}, \Omega\}$

$f(P_3) = \{0, 0, 0, 0, 2, 2, 1, 2, 3\}$

$P_3$ est une pyramide qui n'est ni indicée au sens large ni indicée au sens strict car il n'existe pas deux paliers différents de $\{ab\}$ tels que leur intersection est $\{ab\}$. (Voir figure $(c)$).




*(a)* *(b)* *(c)*

**Définition 4.1.3**

Une pyramide **binaire** est une pyramide dont chaque palier, non réduit à un singleton, est formé par la réunion de deux paliers distincts.

**Remarque :** Une pyramide peut être indicée au sens large et être binaire (voir par exemple $P_2$ dans l'exemple précédent) et toute pyramide peut être rendue binaire.

### 4.1.2 Indices pyramidaux

**Définition 4.1.4**

Un **indice pyramidal** $s$ est un indice de distance qui vérifie en plus la condition suivante :

$\exists$ un ordre $\theta$ sur $\Omega$ tel que tout triplet $(w, w', w'')$, avec $w'$ compris entre $w$ et $w''$ selon $\theta$, satisfait à l'inégalité suivante :

$$s(w, w'') \geq \max\left(s(w, w'), s(w', w'')\right) \text{ (inégalité pyramidale).}$$

On voit donc (d'après la définition 2.2.2 de compatibilité) que tout ordre qui satisfait à l'inégalité pyramidale est compatible avec $s$.

### 4.1.3 Propriétés des indices pyramidaux

**Proposition 4.1.1** *L'ensemble des ultramétriques est inclus dans l'ensemble des indices pyramidaux.*

**Preuve :** Il faut vérifier que les ultramétriques vérifient bien les deux conditions pour être des indices pyramidaux.

Une ultramétrique étant une distance, la première condition est toujours vérifiée.

D'autre part, par la proposition 2.4.3, on sait que pour une ultramétrique $\delta$ donnée on peut toujours lui associer un ordre $\theta$ (non unique) tel que $M(\delta, \theta)$ soit une matrice ultramétrique (et SDD). De plus, par la proposition 2.4.4, on a que :

$M(\delta, \theta)$ est de Robinson $\Leftrightarrow$ $\delta$ et $\theta$ sont compatibles $\Rightarrow$ inégalité pyramidale.

Donc il existe toujours un ordre $\theta$ tel que la deuxième condition est satisfaite.

□

**Proposition 4.1.2** *Les conditions suivantes sont équivalentes si $s$ est un indice pyramidal :*

1. *$\theta$ est compatible avec $s$.*
2. *$M(s, \theta)$ est Robinson*
3. *Tout couple d'éléments $(w, w')$ compris (au sens large) selon $\theta$ entre deux éléments $w_i$ et $w_j$ est tel que $s(w_i, w_j) \geq s(w, w')$.*

**Preuve :**

*1.* $\Longleftrightarrow$ *2.*

On sait (voir au chapitre 2 à la section 2.3) que $\theta$ est compatible avec un indice de dissimilarité $s$ si et seulement si la matrice $M(s, \theta)$ est Robinson.

*3.* $\Longrightarrow$ *1.*

Pour vérifier la compatibilité, c'est-à-dire l'inégalité suivante :

$$\forall (w, w', w_j) \quad \text{ordonné selon } \theta, \quad s(w, w_j) \geq \max\left(s(w, w'), s(w', w_j)\right),$$

il suffit de prendre dans la condition *3.* $w_i = w$ ou $w' = w_j$.
Montrons en prenant $w_i = w$. En effet, on obtient

$$s(w, w_j) \geq s(w, w') \quad \text{avec } (w, w', w_j) \text{ ordonné selon } \theta \ .$$

De plus $(w', w_j)$ est compris (au sens large) entre $w$ et $w_j$ selon $\theta$, donc on a aussi $s(w, w_j) \geq s(w', w_j)$. Dès lors $s(w, w_j) \geq \max(s(w, w'), s(w', w_j))$.

*1.* $\Longrightarrow$ *3.*

En reprenant la définition de compatibilité pour tout triplet quelconque $(w_i, w, w_j)$, on a que

$$s(w_i, w_j) \geq \max\left(s(w_i, w), s(w, w_j)\right).$$

Considérons un $w'$ tel qu'il est compris entre $w$ et $w_j$, alors on a $(w, w', w_j)$ ordonné selon $\theta$. Comme on sait déjà que

$$s(w, w_j) \geq \max(s(w, w'), s(w', w_j)),$$

il suffit de faire le développement suivant

$$\begin{aligned} s(w_i, w_j) &\geq \max\left(s(w_i, w), s(w, w_j)\right) \\ &\geq s(w, w_j) \geq s(w, w') \end{aligned}$$

et on a bien montré que pour tout couple quelconque d'éléments $(w, w')$ compris selon $\theta$ entre $w_i$ et $w_j$, on a que $s(w_i, w_j) \geq s(w, w')$.

□

## 4.2 Existence d'une bijection entre les indices pyramidaux et les pyramides indicées

Il semble utile d'étendre le théorème de bijection entre hiérarchies et ultramétriques (proposition 2.4.1) à l'existence d'une bijection entre pyramides et indices pyramidaux. La preuve de cette bijection est utile car le résultat est important.

**Proposition 4.2.1** *Il existe une bijection entre l'ensemble des pyramides indicées au sens large, noté* $\Pi$*,et l'ensemble des indices pyramidaux, noté* $\mathcal{S}$*.*

**Preuve :** Montrons qu'il existe une application $\phi$ de $\Pi$ dans $\mathcal{S}$ et une application $\psi$ de $\mathcal{S}$ dans $\Pi$ et pouis montrons que $\phi$ et $\psi$ sont inverses l'une de l'autre.

Cette preuve va se faire par construction, c'est-à-dire que l'on montre qu'on peut toujours construire ces deux applications.

**Construction d'une application $\phi$ de $\Pi$ dans $\mathcal{S}$.**

*Soit* $\phi : \begin{array}{ccc} \Pi & \longrightarrow & \mathcal{S} \\ (P,f) & \longmapsto & s \end{array}$ *telle que* $\phi((P,f)) = s$

*avec* $s(k,l) = \inf\left(f(h) \mid h \in P, (k,l) \in h \times h\right)$,
*où* $(P,f)$ *est une pyramide notée* $P$ *indicée au sens large par* $f$.

- **Montrons que $s$ tel qu'il est construit est un indice pyramidal.**

Cela revient à montrer que $s$ vérifie les trois conditions de la définition d'un indice pyramidal :

1. Vérifions la symétrie, c'est-à-dire : $s(k,l) = s(l,k)$, en effet si $(k,l) \in h \times h \Rightarrow (l,k) \in h \times h$ donc

   $$s(k,l) = \inf(f(h) \mid (k,l) \in h \times h) = \inf(f(h) \mid (l,k) \in h \times h) = s(l,k)\ .$$

2. Vérifions maintenant la condition suivante tel que l'indice devient une distance : $s(k,l) = 0 \Leftrightarrow k = l$.

   En effet :

   - $k = l \Rightarrow s(k,l) = 0$ car le $h$ tel que $s(l,l) = f(h)$ est le singleton $h = \{l\}$ d'où $s(l,l) = 0$ par définition de $f$.
   - $s(k,l) = 0 \Rightarrow k = l$ car cela implique que le plus bas palier contenant $k$ et $l$ est de hauteur nulle. Donc ce palier ne peut contenir qu'un seul élément par définition de $f$, donc $k = l$.

3. Vérifions enfin que $s$ satisfait à l'inégalité pyramidale et donc qu'il existe un ordre compatible avec lui. Soit $\theta$ un ordre compatible avec la pyramide $P$, nous allons montrer que $\theta$ est compatible avec $s$.

Nous considérons un triplet quelconque $(w_i, w_j, w_l)$ de $\Omega$ tel que $w_j$ soit compris entre $w_i$ et $w_l$ selon $\theta$. Notons $h_{il}$ le plus bas palier contenant $w_i$ et $w_l$ et notons $h_{ij}$ le plus bas palier qui contient $w_i$ et $w_j$. Comme $h_{il}$ est connexe, il contient l'élément $w_j$ et puisque $P$ est indicée au sens large par $f$, $h_{ij}$ est au maximum à la hauteur de $h_{il}$, donc $s(w_i, w_j) \leq s(w_i, w_l)$. Par le même raisonnement avec le palier $h_{jl}$ étant le plus bas palier contenant $w_j$ et $w_l$, on obtient que $s(w_j, w_l) \leq s(w_i, w_l)$. Finalement, on a que $s(i, l) \geq \max(s(w_i, w_j), s(w_j, w_l))$, donc $\theta$ est bien compatible avec $s$ et $s$ est un indice pyramidal.

**Construction d'une application $\psi$ de $\mathcal{S}$ dans $\Pi$.**

*Soit* $\psi : \begin{array}{ccc} \mathcal{S} & \longrightarrow & \Pi \\ s & \longmapsto & (P, f) \end{array}$ *telle que* $\psi(s) = (P, f)$

*où*

- *$P$ est l'ensemble des parties $h$ de $\Omega$ qui satisfont à la condition :*

$$\begin{cases} \exists \alpha : h = \{x \in \Omega \mid \forall y \in h,\ s(x, y) \leq \alpha\} & (1) \\ \textbf{ou} \\ \exists h_1 \text{ et } h_2 \text{ d'intersection non vide dans } P \text{ tels que} \\ h \neq h_1,\ h \neq h_2 \text{ et } h = h_1 \cap h_2 & (2) \end{cases}$$

- *On note $P(\alpha)$, l'ensemble des parties $h$ de $\Omega$ qui satisfont à la condition (1), pour un $\alpha$ donné.*

- *$f$ est l'application de* $P \to \mathbb{R}^+$ :

$$f(h) = \min(\alpha \mid h \in P(\alpha)).$$

Remarquons que $P(\alpha)$ est en général un recouvrement et non une partition car la relation $s(x, y) \leq \alpha$ entre $x$ et $y$, bien qu'elle soit réflexive et symétrique, n'est pas transitive. En effet, pour un $\alpha$ donné, si $s(x, y) \leq \alpha$ et $s(y, z) \leq \alpha$, ça n'implique pas pour autant que $s(x, z) \leq \alpha$. Ainsi un tel $y$ peut appartenir à deux parties de $P(\alpha)$, une avec $x$ et une avec $z$, et on obtient un recouvrement et non une partition car $y$ appartient aux deux.

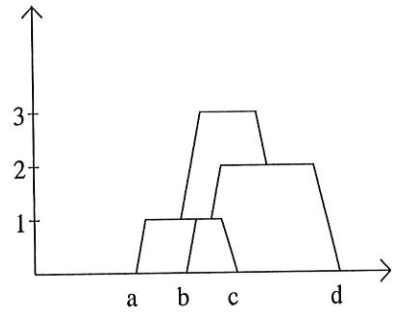
En notant $P'$ l'ensemble formé par tous les éléments de tous les $P(\alpha)$, $\alpha \in \mathbb{R}^+$, il faut également remarquer que ce $P'$ est inclus et non nécessairement identique à $P$ car certaines intersections de classes peuvent ne pas

appartenir à $P'$ tandis qu'elles appartiennent à $P$. Donc $P'$ n'est pas toujours une pyramide et la condition (2) est là pour résoudre ce problème.

Un exemple est donné afin d'illustrer ce qui vient d'être dit sur $P'$.

**Exemple :**

$\Omega = \{a, b, c, d\}$
$P = \{\{a\}, \{b\}, \{c\}, \{d\}, \{abc\}, \{bc\}, \{bcd\}, \Omega\}$
$f(P) = \{0, 0, 0, 0, 1, 1, 2, 3\}$




Comme $\{bc\} \subset \{abc\} \Rightarrow \{bc\} \notin P'$ car pour $\alpha = 1$ c'est $\{abc\}$ qui satisfait à la condition (1), tandis que $\{bc\}$ n'est pas détecté par cette contition, donc $P' = P - \{bc\}$

Par contre $\{bc\} = \{abc\} \cap \{bcd\} \Rightarrow \{bc\} \in P$ grâce à la condition (2).

Toutes les intersections non vides d'éléments de $P$ sont dans $P$, $P$ contient bien $\Omega$ et les singletons, c'est donc une pyramide tandis que $P'$ n'en est pas une.

□

Après avoir clairement présenté la fonction $\psi$ et ses implications, montrons maintenant que son image $(P, f)$ est une pyramide indicée au sens large, nous allons commencer en montrant qu'elle est une pyramide.

- **Montrons que $P$ est une pyramide.**

Tout d'abord, citons deux résultats qui seront utiles pour montrer que $P$ est une pyramide. Le premier résultat est connu par construction et il faut démontrer le deuxième.

(1) $f(h) = \min\big(\alpha \mid \alpha \in \mathbb{R}^+,\ h \in P(\alpha)\big)$ autrement dit $h \in P(f(h))$.

(2) $f(h) = \max\big(s(k,l) \mid (k,l) \in h \times h\big)$ :

En effet, soit $f(h) = \alpha_0$ et $s(x,y) = \max\big(s(k,l) \mid (k,l) \in h \times h\big)$ :

$$\begin{array}{rcl} \text{Supposons } s(x,y) < \alpha_0 & \Rightarrow & \exists \alpha\ :\ s(x,y) \leq \alpha < \alpha_0 \\ & \Rightarrow & h \in P(\alpha) \text{ et } \alpha_0 \neq \text{de la plus petite valeur} \\ & & \text{telle que } h \in P(\alpha_0)\ . \\ & \Rightarrow & \text{Contradiction} \\ & \Rightarrow & s(x,y) \geq \alpha_0 \\ \text{Comme } s(x,y) \leq \alpha_0 & \Rightarrow & f(h) = \alpha_0 = s(x,y)\ . \end{array}$$

Vérifions maintenant que $P$ satisfait bien aux quatre conditions qui définissent une pyramide :

1. $\Omega \in P$ :
   Soit $\alpha = \max\big(s(x,y) \mid (x,y) \in \Omega \times \Omega\big)$,
   alors la partie $h = \{x \in \Omega \mid \forall y \in h,\ s(x,y) \leq \alpha\} = \Omega$ est par définition un élément de $P$.

2. $\forall w \in \Omega,\ \{w\} \in P$ :
   Soit $\alpha = 0$,
   alors la partie $h = \{w \in \Omega \mid \forall y \in h,\ s(w,y) = 0\}$ appartient à $P$
   or comme $s(w,y) = 0 \Leftrightarrow w \equiv y$ par définition de $s$, $h = \{w\} \in P$.

3. $\forall (h_1, h_2) \in P^2,\ h = h_1 \cap h_2 \neq \emptyset \Rightarrow h \in P$ :

   $$\left.\begin{array}{r} \text{Si } h = h_1 \cap h_2 \text{ satisfait à la condition (1)} \\ \text{ou si } h = h_1 \\ \text{ou si } h = h_2 \end{array}\right\} \Longrightarrow h \in P\ .$$

   Si aucune de ces 3 conditions précédentes n'est vérifiée, alors $h \in P$ quand même grâce à la condition (2).

4. Il existe un ordre $\theta$ compatible avec $P$ :

   Soit $\theta$ un ordre compatible avec $s$, montrons que cet ordre est compatible avec $P$.

   Prenons $h \in P$ et montrons que $h$ est connexe selon $\theta$, c'est-à-dire que si $w'$ et $w''$ sont les bornes de $h$ selon $\theta$, on a :

   (i) $\{w \in h\} \Rightarrow \{w$ compris entre $w'$ et $w''$ selon $\theta\}$ ,
   ce qui revient à montrer que
   $\{w \notin [w', w''] \text{ selon } \theta\} \Rightarrow \{w \notin h\}$ .

   (ii) $\{w$ compris entre $w'$ et $w''$ selon $\theta\} \Rightarrow \{w \in h\}$

   Montrons (i) :
   Soit $w \notin [w', w'']$ selon $\theta$, alors :
   soit $w$ est à gauche de $w'$ et $s(w, w'') > s(w', w'') = f(h)$ car $s$ est compatible avec $\theta$,
   soit $w$ est à droite de $w''$ et $s(w', w) > s(w', w'') = f(h)$.
   Donc dans les deux cas $w \notin h$.

   Montrons (ii) :
   Soit $w' \leq w \leq w''$ selon $\theta$.
   $\forall w^* \in h$, $w^* \in [w', w'']$ selon $\theta$ par (i), donc d'après la proposition 4.1.2 :

$$\begin{aligned} \forall w^* \in h,\ s(w, w^*) &\leq s(w', w'') \\ &= \max\left(s(k, l) \mid (k, l) \in h \times h\right) \\ &= f(h) \quad \text{(par (2))} \\ \Rightarrow s(w, w^*) &\leq f(h). \end{aligned}$$

   Or $h = \{x \in \Omega \mid \forall w^* \in h,\ s(x, w^*) \leq f(h)\} \Rightarrow w \in h$.

On a donc bien montré que $P$ est une pyramide car elle vérifie les quatre conditions.

• **Montrons que $(P, f)$ est une pyramide indicée au sens large.**

Pour cela, montrons que $P$ vérifie bien les trois conditions de la définition :

1. $f(h) = 0 \Leftrightarrow h$ est un singleton :
   En effet :

$$\begin{aligned} f(h) = 0 &\Leftrightarrow \min(\alpha \mid h \in P(\alpha)) = 0 \\ &\Leftrightarrow h \in P(0) \\ &\Leftrightarrow \forall y \in h,\ s(x, y) = 0 \\ &\Leftrightarrow \forall y \in h,\ x \equiv y \\ &\Leftrightarrow h \text{ ne contient qu'un seul élément.} \end{aligned}$$

2. $\forall (h, h') \in P \times P, h \subset h'$ (strictement) $\Leftrightarrow f(h) \leq f(h')$ :
   Soit $w' \in h^c \cap h'$, alors pour au moins un $w \in h$, on a $s(w, w') \geq f(h)$, sinon $w' \in h$.
   D'autre part,
   $s(w, w') \leq f(h') = \max(s(k, l) \mid (k, l) \in h' \times h')$ car $(w, w') \in h' \times h'$
   d'où $f(h) \leq s(w, w') \leq f(h')$, on a bien l'inégalité recherchée.

3. Il faut montrer que :

$$\left.\begin{array}{l} h \subset h' \text{ (strictement)} \\ f(h) = f(h') \end{array}\right\} \Longrightarrow \exists\ h_1 \neq h \text{ et } h_2 \neq h \in P : h = h_1 \cap h_2\ .$$

   Comme $h \neq$ du plus grand palier de hauteur $f(h)$, il ne peut satisfaire à la condition (1) de la construction de la fonction.
   Donc pour que $h \in P$, il doit nécessairement satisfaire à la condition (2) qui correspond à ce que l'on veut montrer.

Ainsi nous avons enfin montré que les deux applications sont bien contruites.

Maintenant, afin de montrer qu'il existe bien une bijection entre l'ensemble des pyramides indicées au sens large et l'ensemble des indices pyramidaux, il faut montrer le résultat qui suit.

**Les applications $\phi$ et $\psi$ sont inverses l'une de l'autre.**

Il faut démontrer premièrement que $\phi \circ \psi(s) = s$ et deuxièmement que $\psi \circ \phi((P, f)) = (P, f)$.

- **Montrons que $\phi \circ \psi(s) = s$.**

Autrement dit, il faut montrer que :

$$\left.\begin{array}{l} \phi((P,f)) = \sigma \\ \psi(s) = (P,f) \end{array}\right\} \Rightarrow \sigma = s \ .$$

Soit $x, y \in \Omega$ quelconques et soit $h$ un palier de plus petite hauteur qui les contient.
Par définition de $\sigma$ (car obtenu par $\phi$), on sait que

$$\sigma(x,y) = \inf\left(f(h') \mid h' \in P,\ (x,y) \in h' \times h'\right) = f(h) \ .$$

Soit $f(h) = \alpha$, or par la définition de $\psi$,

$$\left.\begin{array}{l} h = \{x \in \Omega \mid \forall y \in h,\ s(x,y) \leq f(h)\} \\ \\ \text{tel que } f(h) = \min\left(\alpha' \mid h \in P(\alpha')\right) \end{array}\right\} \Rightarrow s(x,y) \leq \alpha$$

D'autre part $s(x,y) \not< \alpha$, sinon

$$\exists h' \ni x, y \ :\ f(h') < f(h) \Rightarrow \text{ contradiction avec la définition de } h.$$

D'où $s(x,y) = \alpha = \sigma(x,y)$. Ce résultat pouvant être montré $\forall x, y \in \Omega$, on a que $\sigma = s$ et donc $\phi \circ \psi(s) = s$.

- **Montrons que $\psi \circ \phi((P, f)) = (P, f)$.**

Cela revient à montrer que :

$$\left.\begin{array}{l} \phi((P,f)) = \sigma \\ \psi(\sigma) = (P',f') \end{array}\right\} \Rightarrow P \equiv P' \text{ et } f \equiv f' \ .$$

1. $?P \equiv P'$ ?

Afin de montrer l'identité entre $P$ et $P'$, on va démontrer la suite d'équivalences suivante :

$$\{h \in P\} \quad \textbf{(a)}$$

$$\Leftrightarrow \left\{ \begin{array}{l} \{\exists (i,j) \in h \times h \mid h = \{x \in \Omega \mid \forall y \in h, \sigma(x,y) \leq \sigma(i,j)\}\} \textbf{(1)} \\ \\ \text{ou } \{\exists h' \text{ et } h'' \text{ dans } P \text{ avec } h' \neq h \text{ et } h'' \neq h : h = h' \cap h''\} \textbf{(2)} \end{array} \right\} \textbf{(b)}$$

$$\Leftrightarrow \quad \{h \in P'\} \quad \textbf{(c)}$$

**(a) $\Rightarrow$ (b) :**

$\phi((P,f)) = \sigma$ et $h \in P$ impliquent l'existence de $i$ et $j$ dans $h$ tels que $\forall (x,y) \in h \times h$, $\sigma(x,y) \leq \sigma(i,j)$. On peut choisir $i$ et $j$ de façon qu'ils constituent les éléments extrêmes de la partie connexe associée à $h$ selon un ordre $\theta$ compatible avec $\sigma$.

Puisque $\sigma$ est pyramidal, par la proposition 4.1.2, on a bien $\sigma(i,j) \geq \sigma(x,y)$, $\forall x, y$ compris entre $i$ et $j$ selon l'ordre $\theta$. Si $h$ contient tous les éléments $w$ de $\Omega$ tels que $\forall y \in h, \sigma(w,y) \leq \sigma(i,j)$ alors la condition **(1)** est satisfaite.

Sinon, soit $w \notin h$ tel que $\sigma(w,y) \leq \sigma(i,j), \forall y \in h$. Si $i$ se situe entre $w$ et $j$ au sens de $\theta$, on a $\sigma(w,j) \geq \sigma(i,j)$ puisque $\sigma$ est pyramidal. Or, par définition de $w$, on a $\sigma(w,j) \leq \sigma(i,j)$ d'où $\sigma(w,j) = \sigma(i,j)$, (on fait bien sur le même raisonnement si c'est $j$ qui se situe entre $w$ et $i$. Donc le palier $h'$ de plus basse hauteur qui contient $w$ et $j$ est à la même hauteur que $h$. Comme il contient $w$ et $j$, il contient tous les éléments intermédiaires, donc ceux compris entre $i$ et $j$, donc $h$. Il en résulte que $h \subset h'$ et que $f(h) = f(h')$.

Comme $(P,f)$ est une pyramide indicée au sens large, par définition, on a bien qu'il existe $h_1$ et $h_2$ distincts dans $P$ tels que $h \neq h_1$, $h \neq h_2$ et $h = h_1 \cap h_2$.

**(b) $\Rightarrow$ (a) :**

Montrons que si la condition **(1)** ou **(2)** est satisfaite, alors $h \in P$.

Si **(1)** est vrai : soit $h'$ un palier de $P$ de plus basse hauteur qui contient $i$ et $j$ (les 2 points les plus éloignés de $h$ selon $\theta$), donc $f(h') = \sigma(i,j)$

et $h \subset h'$. De plus, $h' \subset h$ car

$$\forall w \notin h,\ \exists y \in h : \sigma(w,y) > \sigma(i,j),$$

donc $w$ ne peut appartenir à $h'$ puisque $f(h') = \sigma(i,j)$, on a donc que :

$$\left.\begin{array}{l} h' \subset h \\ h \subset h' \end{array}\right\} \Rightarrow h' \equiv h \Rightarrow h \in P\ .$$

Si **(2)** est vrai : $h \in P$ par définition d'une pyramide.

**(b) $\Leftrightarrow$ (c) :**
Soit $(P', f')$ la pyramide obtenue par $\psi(\sigma)$, par définition de $P'$, on voit immédiatemment par l'équivalence suivante que la condition **(b)** est équivalente à la définition des éléments de $P'$ par définition des $P'(\alpha)$ :

$$\textbf{(b)} \Leftrightarrow h = \{x \in \Omega \mid (i,j) \in h \times h : \forall y \in h,\ \sigma(x,y) \leq \sigma(i,j)\}$$

$$\Leftrightarrow h = \{x \in \Omega \mid \forall y \in h,\ \sigma(x,y) \leq \alpha\} \quad \text{si } \alpha = \sigma(i,j) \Leftrightarrow \textbf{(c)}$$

2. $?f \equiv f'\,?$
Autrement dit, il faut montrer que $f'(h) = f(h),\ \forall h \in P$.
En effet, $\forall h \in P' \equiv P$, on a que

$$f'(h) = \min(\alpha \in \mathbb{R}^+ \mid h \in P(\alpha)) \Rightarrow f'(h) = \sigma(i,j)$$

où $\sigma(i,j) = \max(\sigma(l,k) \mid (l,k) \in h \times h)$.

D'autre part, par définition de $\phi$,

$$\sigma(i,j) = \inf(f(h) \mid h \in P,\ (i,j) \in h \times h)\ .$$

Comme $i$ et $j \in h$, on a que $\sigma(i,j) \leq f(h)$.
Soient $x$ et $y$, les deux éléments de $h$ les plus éloignés selon l'ordre $\theta$, comme $\sigma$ est pyramidal, on a $\sigma(x,y) \geq \sigma(i,j)$.
Or par définition de $\sigma(i,j)$, on a $\sigma(i,j) \geq \sigma(x,y)$, d'où $\sigma(i,j) = \sigma(x,y)$.

Montrons que $\sigma(x,y) = f(h)$.

Si on avait $\sigma(x,y) \leq f(h)$, par définition de $\sigma$ qui est obtenu par l'application $\phi$, on aurait

$$\exists h' \text{ plus bas que } h : \sigma(x,y) = f(h') \Rightarrow f(h') \leq f(h) \ .$$

D'autre part, par définition de $\theta$, on aurait que $h \subset h'$. En effet, $h$ et $h'$ contiennent tous les éléments compris entre $x$ et $y$ selon $\theta$ et $h$ ne contient que ceux-ci, donc $f(h) \leq f(h')$ car $P$ est une pyramide indicée au sens large.

On obtient donc bien que $f(h) = f(h') = \sigma(x,y) = \sigma(i,j)$. Comme on peut faire ce raisonnement $\forall h \in P$, on a bien montré que

$$\forall h \in P,\ f(h) = \sigma(i,j) = f'(h) \ .$$

Ainsi se termine la démonstration du théorème de bijection, bien qu'elle soit très longue, elle est utile car ce résultat est important. En effet, on peut maintenant affirmer que pour une pyramide indicée au sens large, on peut toujours lui associer un indice pyramidal et inversement.

□

En reprenant la définition 3.1.4 d'un ordre donnant lieu à un croisement, la proposition 4.1.2 et les résultats obtenus tout au long de la démonstration de la proposition 4.2.1 précédente, le résultat suivant est obtenu facilement en ayant $\phi(P) = s$.

**Proposition 4.2.2** *Les propriétés suivantes sont équivalentes :*

- *$\theta$ est sans croisement pour la pyramide $P$.*
- *$\theta$ est compatible avec $P$.*
- *$\theta$ est compatible avec $s$.*
- *$M(s,\theta)$ est Robinson.*

# Chapitre 5

# Hiérarchies et pyramides

## 5.1 Hiérarchies et pyramides saturées

**Définition 5.1.1**

Une hiérarchie ou une pyramide est **saturée** quand le nombre de ses paliers est maximum.

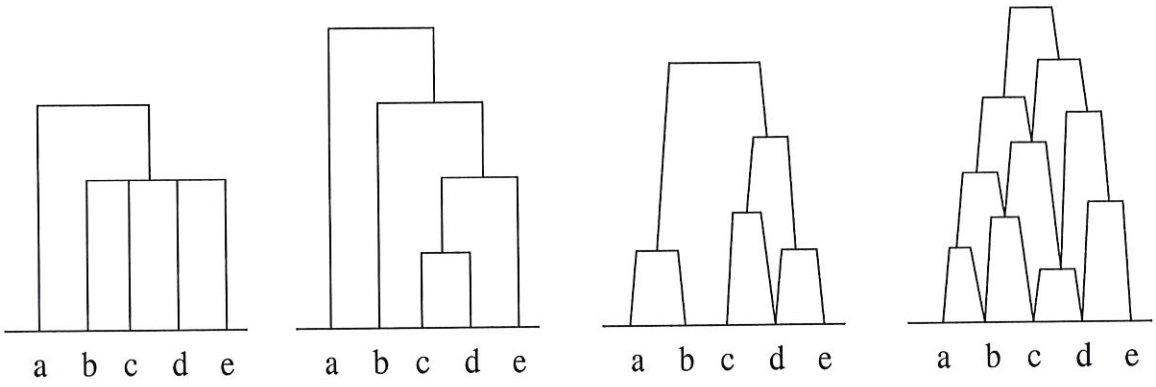
A partir de cette définition, on constate qu'une hiérarchie sur $n$ objets est saturée quand le nombre de ses paliers est égal à $n-1$.

Par contre, une pyramide sur $n$ objets est saturée quand le nombre de ses paliers est égal au nombre maximum de distances différentes possibles entre les objets, c'est-à-dire $\frac{n(n-1)}{2}$.

Il y a donc, dans une pyramide saturée, $\frac{n}{2}$ fois plus de paliers que dans une hiérarchie saturée.

Par exemple, pour $n = 200$, il y a 100 fois plus de paliers dans la pyramide saturée que dans la hiérarchie saturée et l'utilisateur va avoir beaucoup de mal à interpréter une pyramide à $\frac{200 \times 199}{2} = 19.900$ paliers. Une telle constatation est un inconvénient pour les pyramides et afin d'éviter cette difficulté, on est conduit à chercher des pyramides non saturées.

La figure suivante illustre des exemples de hiérarchies et pyramides saturées et non saturées.




Hiérarchie non saturée | Hiérarchie saturée | Pyramide non saturée | Pyramide saturée

## 5.2 Construction de pyramides non saturées

Afin de construire des pyramides non saturées, pour que les résultats d'une classification pyramidale soient interprétables, on peut avoir recours à deux stratégies différentes.

Premièrement, une solution est de partir d'une hiérarchie pour l'enrichir de paliers qui la rende pyramidale sans créer d'inversions, c'est-à-dire qu'aucun palier n'est plus haut qu'un palier qui le contient.

Deuxièmement, on peut partir d'une pyramide saturée et supprimer des paliers inutiles quand il y en a qui sont trop voisins. Ces deux méthodes sont détaillées ci-dessous.

### 5.2.1 Pyramidisation d'une hiérarchie

Cette pyramidisation peut se faire en trois étapes :

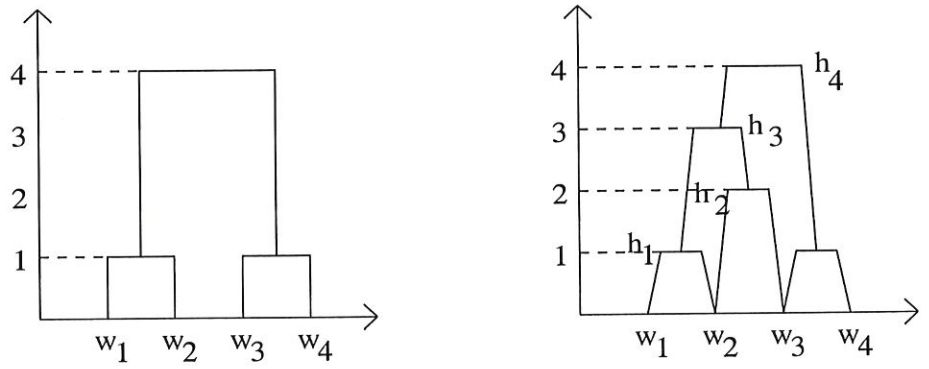
1. Construire une hiérarchie à l'aide d'un indice d'agrégation $\delta$.
2. Choisir un ordre $\theta$ sans croisement (définition 3.1.4) pour cette hiérarchie.
3. Considérer toutes les classes consécutives selon $\theta$ en commencant par les plus basses. Les réunir, afin de former un nouveau palier, chaque fois que la hauteur de ce nouveau palier est inférieure à la hauteur du plus bas palier de la hiérarchie qui le contient.

Si les paliers ainsi obtenus sont trop nombreux et forment une pyramide trop chargée, on utilise alors la deuxième stratégie devellopée dans la section 5.2.2 suivante.

**Exemple :**
Considérons la matrice de dissimilarité suivante :

$$\begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \end{array} \begin{array}{c} \begin{array}{cccc} w_1 & w_2 & w_3 & w_4 \end{array} \\ \begin{pmatrix} 0 & 1 & 3 & 4 \\ & 0 & 2 & 4 \\ & & 0 & 1 \\ & & & 0 \end{pmatrix} \end{array}$$

Si on choisit l'indice d'agrégation de la distance maximum, on obtient la hiérarchie de la première des figures suivantes et la seconde figure est la pyramide résultant de la pyramidisation de cette hiérarchie.




Hiérarchie de départ

Pyramidisation de la hiérarchie

FIG. 5.1 – Pyramidisation d'une hiérarchie.

La pyramide qui est ainsi obtenue contient deux paliers de plus que la hiérarchie de départ. En effet, les paliers $h_2$ et $h_3$ ont pû apparaître successivement car leur hauteur était plus basse que celle du plus bas palier qui les contenait, à savoir le palier $h_4$.

### 5.2.2 Hiérarchisation d'une pyramide

On part d'une pyramide, celle-ci induit un indice pyramidal et donc une matrice de dissimilarité $M(s, \theta)$ où $\theta$ est un ordre associé à $s$.

En utilisant $s$, on peut rendre chaque triangle $(w_i, w_j, w_k)$ isocèle avec la base plus petite que les côtés en utilisant un nouvel indice de dissimilarité $s'$ tel que

$$s'(w_i, w_k) = \max(s(w_i, w_j), s(w_j, w_k)),$$

si $s(w_i, w_k)$ n'est pas le plus petit côté du triangle $(w_i, w_j, w_k)$, (au lieu de max, on pourrait aussi prendre le min).

Ainsi en reprenant l'exemple de la section 5.2.1 précédente avec la même matrice de dissimilarité , on transforme $(w_1, w_2, w_3)$ de la façon suivante :

$$\begin{aligned} s(w_1, w_2) = 1 &\longrightarrow s'(w_1, w_2) = 1 \ , \\ s(w_2, w_3) = 2 &\longrightarrow s'(w_2, w_3) = 3 \ \text{ et} \\ s(w_1, w_3) = 3 &\longrightarrow s'(w_1, w_3) = 3 \ . \end{aligned}$$

Chaque fois qu'un triangle est rendu isocèle, on se rapproche d'une hiérarchie. À la limite quand les $C_n^3 = \frac{n!}{3!(n-3)!}$ triangles sont rendus isocèles, avec la base plus petite que les côtés, l'indice $s'$ est une ultramétrique.

En pratique, on peut utiliser un algorithme de Hiérarchisation d'une Pyramide (HDP) mais il n'est pas utile de développer cet aspect d'autant plus que le programme de classification pyramidale, présenté dans la partie pratique de ce travail, ne permet pas d'obtenir une hiérarchisation.

**Exemple 5.2.1**
Considérons la pyramide saturée donnée par le premier dessin de la figure 5.2 et la matrice de Robinson suivante qui lui est associée.

$$\begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \end{array} \begin{array}{c} \begin{array}{cccc} w_1 & w_2 & w_3 & w_4 \end{array} \\ \begin{pmatrix} 0 & 1 & 4 & 7 \\ & 0 & 2 & 5 \\ & & 0 & 4 \\ & & & 0 \end{pmatrix} \end{array}$$

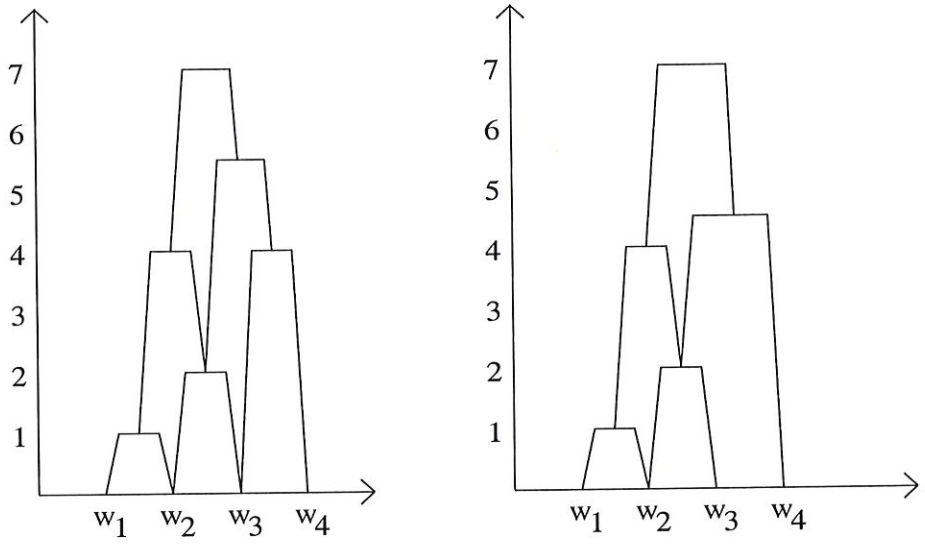



FIG. 5.2 – Hiérarchisation d'une pyramide.

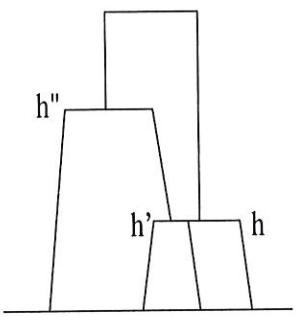
Le résultat de la hiérarchisation de la pyramide est donné par le deuxième dessin de la figure 5.2. La matrice de dissimilarité obtenue est la suivante.

$$\begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \end{array} \begin{array}{c} \begin{array}{cccc} w_1 & w_2 & w_3 & w_4 \end{array} \\ \begin{pmatrix} 0 & 1 & 4 & 7 \\ & 0 & 2 & 4.5 \\ & & 0 & 4.5 \\ & & & 0 \end{pmatrix} \end{array}$$

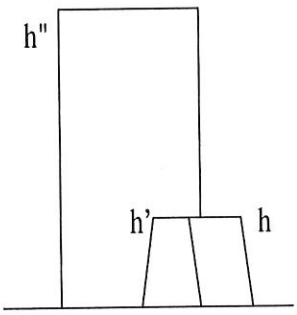
L'égalisation de paliers la plus avantageuse est obtenue dans le triangle $(w_2, w_3, w_4)$ car $s(w_4, w_2) - s(w_4, w_3) = 1$ est le plus petit écart. On obtient de cette manière une nouvelle pyramide où les paliers $(w_2, w_3, w_4)$ et $(w_3, w_4)$ sont réunis en un seul à la hauteur 4.5. L'arête qui reie $w_3$ au palier $(w_2, w_3, w_4)$ devient alors inutile et on la supprime.

## 5.3 Suppression d'arêtes inutiles : Epuration

L'épuration porte d'abord sur les paliers redondants, on élimine tous les paliers qui sont inutiles pour la représentation visuelle de la pyramide. Autrement dit, on supprimera tous les paliers $h \in P$ tels qu'il existe $h' \in P$ avec $f(h) = f(h')$, $h' \subset h$ et qu'il n'existe pas de $h'' \in P$ différent de h et tel que $h' \subset h''$, la figure suivante illustre bien cette règle. De cette façon, on aboutit toujours à une pyramide indicée au sens large.




$h' = h \cap h''$
$\Rightarrow h'$ n'est pas supprimé




$h' \subset h$ et $f(h') = f(h)$
$\Rightarrow h'$ est supprimé

Une autre forme d'épuration concerne les "arêtes" de la pyramide. Une arête est le segment de droite qui relie un palier aux plus petits paliers qui le contiennent (deux au maximum).

En effet, une fois l'épuration des paliers réalisée, on s'aperçoit que certains d'entre eux sont reliés par une arête, à plusieurs paliers qu'ils contiennent. Un exemple est donné par la figure 5.3 où on voit que le palier $h$ est relié par 4 arêtes notées 1, 2, 3, 4 aux paliers $h_1, h_2, h_3, h_4$. Ces arêtes servent à montrer l'inclusion d'un palier dans un autre, par exemple l'arête 2 sert à montrer l'inclusion du palier $h_2$ dans le palier $h$. Les arêtes "extrêmes", comme 1 et 4 dans la figure 5.3, sont indispensables mais certaines arêtes "intérieures",

comme 2 ou 3 dans la figure 5.3, deviennent inutiles si l'inclusion est montrée par un palier intermédiaire. Dans la figure 5.3, l'arête 3 peut être supprimée car elle est redondante avec l'arête 4 qui relie $h_3$ à $h$ par l'intermédiaire du palier $h_4$.

Une règle générale de suppression d'arêtes peut être la suivante : étant donné un palier $h$, on supprime toutes les arêtes "intérieures" qui relient plus d'une fois ce palier avec les paliers qu'il contient.

Finalement pour l'exemple donné par la figure 5.3, les arêtes 1 et 4 qui sont extrêmes doivent être conservées, l'arête 3 doit être supprimée et l'arête 2 qui n'est reliée qu'une fois au palier $h$ doit être conservée.

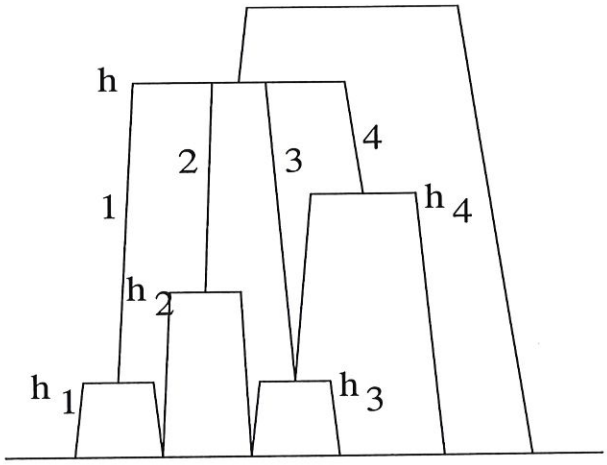



FIG. 5.3 – Pyramide non epurée

**Remarque :** En pratique, l'algorithme présenté dans la deuxième partie de ce travail permettra de faire l'épuration mais beaucoup plus simplement car il suffira de fixé un seuil et toute pyramide de hauteur trop proche de celle d'une classe un peu plus haute sera supprimée.

# Deuxième partie

# Approche Pratique

# Chapitre 6

# Présentation du programme HIPYR

## 6.1 Introduction

Le programme HIPYR est un module de classification symbolique disponible dans le logiciel SODAS 2. Ce programme est en développement constant et a été créé dans le cadre du projet européen ASSO[1]. Ce programme est programmé en C et a été conçu par Paula Brito de l'Université de Porto.

## 6.2 Principes de la méthode

Les informations suivantes sur le programme de classification Hipyr sont tirées des références [6] et [7] de la bibliographie.

Le module Hipyr est une méthode de classification hiérarchique ou pyramidale disponible dans le logiciel SODAS 2, il permet de construire une hiérarchie ou une pyramide à partir d'un ensemble de données symboliques.

Le but général d'une méthode de classification est d'agréger des objets (individus) similaires d'un ensemble de départ $E$ dans des classes homogènes, sur base de valeurs observées pour un ensemble de variables sur les individus. Hipyr suppose donc qu'on dispose d'un ensemble d'entités appelées "individus" (personnes, institutions, cités, objets, etc...) qu'on souhaite organiser en une structure de classes imbriquées, c'est-à-dire une structure d'arbre.

[1] Analysis System of Symbolic Official data

Les deux structures possibles, pour les classes obtenues par Hipyr, sont les hiérarchies ou les pyramides.

Les individus, associés à des objets symboliques, peuvent être décrits par des variables quantitatives simples, intervalles, qualitatives simples, multivaluées catégoriques et/ou modales, des variables avec de différents types sont permises. Si une matrice de dissimilartité est disponible, il est également possible de l'utiliser pour une classification numérique.

La classification hiérarchique ou pyramidale se fait selon une approche ascendante, c'est-à-dire de "bas" en "haut" : les objets les plus similaires sont réunis ensemble, ensuite les classes les plus semblables sont rassemblées, jusqu'à obtenir une classe unique rassemblant tous les éléments de $E$.

Le modèle de classification à utiliser, hiérarchique ou pyramidale, doit être choisi par l'utilisateur. Dans le cas hiérarchique, chaque niveau de la structure correspond à une partition. Dans le cas pyramidal, à chaque niveau on obtient une famille de classes empiétantes, c'est-à-dire un recouvrement et toutes les classes sont des intervalles d'un ordre linéaire total sur $E$. D'où, une pyramide fournit à la fois une classification et une sériation sur les données, (il y a un ordre sur les individus).

Comme le modèle pyramidal conduit vers un système de classes plus riche que celui qui est produit par le modèle hiérarchique, il permet l'identification de classes que le modèle hiérarchique ne pourrait identifier et l'existence d'un ordre compatible sur les objets conduit vers une structure qui est relativement simple.

Pour faire la classification, deux approches sont possibles mais avant de les décrire, introduisons deux définitions nécessaires.

**Définition 6.2.1** (Voir référence [8])

Un objet symbolique est dit **complet**

- s'il est définit par toutes les propriétés qui caractérisent son extension
- s'il est le plus spécifique à remplir cette condition, (c'est-à-dire le moins général).

**Remarque :** L'union préserve la complétude.

**Exemple 6.2.1** *D'un objet symbolique complet.*
Soit la matrice de données suivante :

$$\begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \end{array} \begin{array}{c} \begin{array}{ccc} \textit{Age} & \textit{Poids} & \textit{Sexe} \end{array} \\ \begin{pmatrix} 20 & 45 & F \\ 50 & 55 & F \\ 30 & 50 & F \\ 60 & 60 & M \end{pmatrix} \end{array}$$

Un exemple d'objet symbolique non complet est l'objet $a$ suivant :

$$a = [\text{Poids} = [40, 50]] \wedge [\text{Age} = [20, 50]] \ .$$

Son extension est telle que $ext(a) = \{w_1, w_3\}$ et $a$ n'est **pas complet.**

Un exemple d'objet symbolique complet est l'objet $a'$ suivant :

$$a' = [\text{Poids} = [45, 50]] \wedge [\text{Age} = [20, 30]] \wedge [\text{Sexe}=\{\text{F}\}] \ .$$

Son extension est aussi telle que $ext(a) = \{w_1, w_3\}$ et maintenant $a$ est **complet.**

En effet, les deux objets symboliques $a$ et $a'$ ont la même extension et contiennent tous les deux les propriétés caractérisant leurs éléments mais $a'$ est moins général que $a$ tout en contenant une information suffisante et pas inutile, son degré de généralité sera moins grand.

**Définition 6.2.2** (Voir référence [4])

Soit $E$ l'ensemble des individus et soit $s$ un objet symbolique complet représentant la classe $p$ tel que son extension sur $E$ vaut $ext(s|E) = p$.
Alors, la paire $(p, s)$ est appellée un **concept.**

Décrivons maintenant les deux approches de classification possibles :

1. La classification est basée sur l'ensemble des données symboliques calculées sur des objets.

   Dans ce cas, chaque classe formée est, par construction, associée à un objet symbolique qui est une conjonction des propriétés sur les données symboliques des objets de départ. Ce représentant de la classe constitue une condition nécessaire et suffisante pour l'appartenance aux classes, c'est-à-dire qu'il généralise les membres de la classe qu'il représente et

aucun élément en dehors de la classe ne rencontre la description donnée par cet objet symbolique. Les classes sont désormais des "concepts" décrites à la fois "en extension", par l'ensemble de leurs membres, et "en intention", par cet objet symbolique qui exprime leurs propriétés.

A chaque pas, la méthode définit un critère numérique supplémentaire qui permet de choisir la "meilleure" agrégation parmi les agrégations possibles. Ce critère est le **degré de généralité** qui, pour des variables intervalles et catégoriques multivaluées, évalue la proportion du domaine sous-jacent qui est couvert par un objet symbolique représentant une classe. Et pour des variables modales, ce critère évalue de combien la distribution donnée pour une classe est proche d'une distribution uniforme. Davantage de détails au sujet de ce degré de généralité seront donnés en section 6.5.2.
Chaque fois qu'une classe est formée, deux mesures sont disponibles : la valeur du degré de généralité de la classe formée et l'augmentation du degré de généralité résultant de la formation de la classe.

2. La classification est basée sur une matrice de dissimilarité entre les éléments de $E$.
Dans ce cas, ces dissimilarités doivent être calculées par le module DISS et la donnée d'entrée de HIPYR est un fichier contenant une matrice de dissimilarité triangulaire. Ensuite, un algorithme classique de classification hiérarchique ou pyramidale est appliqué, c'est-à-dire qu'à chaque pas, les classes les plus similaires sont agrégées. Les différentes mesures d'agrégation, qui évaluent les dissimilarités entre les classes, sont considérées, telles que le lien complet, le lien moyen, le lien simple ou le diamètre.
Dans ce cas, les classes ne sont pas automatiquement associées à une description discriminante.

Dans le premier cas, la description de chaque classe est donnée par l'objet symbolique associé, chaque classe est définie par l'ensemble de ses membres et par sa description qui est un objet symbolique qui généralise ses membres. Dans le second cas, seules les valeurs des dissimilarités sont disponibles, ainsi aucune description n'est donnée pour les classes.

Une fois que la structure est terminée, une mesure de comparaison entre les individus peut être obtenue, c'est une mesure de dissimilarité induite.

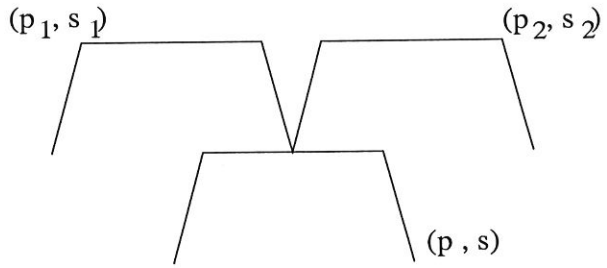
Pour n'importe quel couple d'individus, la dissimilarité induite est égale à la hauteur de la classe où ils apparaissent ensemble pour la première fois quand on explore la structure de bas en haut. Cette dissimilarité induite peut alors être comparée avec les valeurs de la dissimilarité de départ ou du degré de généralité obtenu directement à partir des données.

Si la hiérarchie ou la pyramide est construite à partir d'une table de données symboliques, nous obtenons une structure d'héritage, dans le sens où chaque classe hérite des propriétés associées à ses successeurs.

Cela va permettre de générer des règles entre les classes, deux méthodes sont considérées :

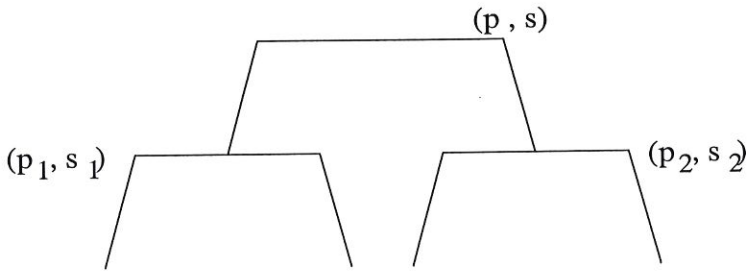
- Méthode de fusion (seulement pour les pyramides) :
  Soient $(p_1, s_1)$ et $(p_2, s_2)$ deux concepts de classes dans la pyramide, et soit $(p, s)$ un autre concept tel que $p = p_1 \cap p_2$. Nous pouvons alors écrire la règle suivante :

$$s_1 \wedge s_2 \Rightarrow s \ .$$




- Méthode de fission (pour les hiérarchies et les pyramides) :
  Soit $(p, s)$ un concept obtenu par la fusion de $(p_1, s_1)$ avec $(p_2, s_2)$, alors $p = p_1 \cup p_2$ et $s = s_1 \cup s_2$. Nous avons alors la règle suivante :

$$s \Rightarrow s_1 \vee s_2 \ .$$




Les visualisations graphiques des pyramides ou des hiérarchies sont exécutées par le module VPYR dans le logiciel SODAS 2.

## 6.3 Epuration

L'épuration d'une hiérarchie ou d'une pyramide consiste à supprimer des classes de la hiérarchie ou de la pyramide dans le but d'obtenir une structure qui est plus facile à interpréter, sans que la perte d'information ne soit importante.

Soit une classe $C$ avec un seul prédécesseur $C'$ (car dans le cas des pyramides une classe peut avoir jusqu'à deux prédécesseurs). Soit $f$ la fonction d'indiçage. Supposons que $f(C') - f(C) < \epsilon$ où $\epsilon > 0$ est un seuil fixé, alors nous pouvons supprimer $C$ de la structure, sans une grande perte. Le choix de $\epsilon$ dépend du degré de simplification que nous souhaitons atteindre, un $\epsilon$ trop petit ne changera presque pas la structure mais un $\epsilon$ grand la simplifiera trop fort et enlèvera beaucoup de classes. Habituellement, $\epsilon$ sera un pourcentage convenable de la hauteur maximum. Dans HIPYR, l'épuration peut être exécutée à partir de la représentation graphique. Une fois que cette représentation est affichée, l'utilisateur a la liberté de demander une simplification de la structure en choisissant le paramètre d'épuration $\epsilon$ qu'il désire. Cette option d'épuration est disponible dans la barre d'outils "Options".

## 6.4 Sortie de HIPYR : données et représentations graphiques

L'algorithme de HIPYR produit en sortie un fichier .sds contenant à la fois les données de départ et un nouvel ensemble d'objets symboliques : les classes de la hiérarchie ou de la pyramide. Des résultats supplémentaires sont fournis dans un fichier texte tels que :

- Fichier et options utilisés, (le critère et la base de données choisis).
- Liste des classes et pour chaque classe :
  - les classes regoupées ensemble pour former la classe présente
  - la liste de ses membres
  - la valeur de l'indice de la classe (sa hauteur)
  - l'objet symbolique représentant la classe si la hiérarchie ou la pyramide est construite par une méthode de classification symbolique. Alors, cette description symbolique constitue une condition nécessaire et suffisante pour l'appartenance à la classe.

- La valeur d'évaluation qui évalue l'ajustement entre la structure obtenue et les données originales, au plus bas est sa valeur et au mieux est l'ajustement.
- La matrice de dissimilarité induite, si elle est demandée.

Une sortie graphique est également fournie par VPYR, plusieurs options sont ensuite disponibles pour explorer la structure :

- Une classe est selectionnée en cliquant dessus. Après, l'utilisateur peut obtenir la description de la classe en termes d'une liste de variables choisies ou de sa représentation par un "Zoom Star".
- L'utilisateur peut épurer la hiérarchie ou la pyramide en utilisant les hauteurs d'agrégations comme un critère. Premièrement, le taux de simplification doit être sélectionné, ensuite l'utilisateur doit cliquer sur le bouton de simplification dans le menu et ceci ouvre une nouvelle fenêtre graphique dans laquelle se trouve un graphique de la pyramide simplifiée.
- Si la hiérarchie ou la pyramide est construite à partir d'une table de données symboliques, les règles, (de fusion et fission), peuvent être générées et sauvées dans un fichier spécifié.
- Si l'utilisateur est intéressé par une classe en particulier, il peut obtenir une fenêtre avec la structure restreinte à cette classe et à ses successeurs.

La figure 6.1 représente la fenêtre qu'on obtient sur SODAS 2 lorsque la classification est terminée. Il faut cliquer sur le symbole représentant un texte pour obtenir le fichier texte et cliquer sur celui représentant un graphe pour obtenir le graphique produit par VPYR.

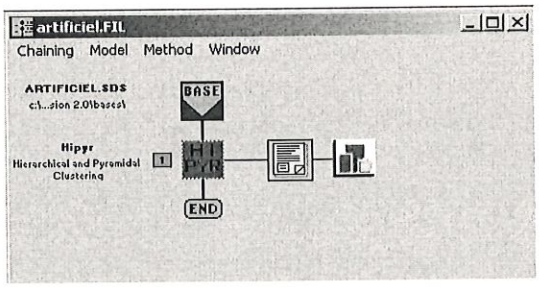



FIG. 6.1 – Fenêtre obtenue quand HIPYR a fini de tourner.

## 6.5 Fonctionnement d'HIPYR

### 6.5.1 Classification classique

Soit $E = \{w_1, \ldots, w_n\}$ un ensemble d'unités (individus, groupes, ...) que nous souhaitons classer. Dans cette situation où la classification se base sur une matrice de dissimilarité, l'algorithme classique de la méthode hiérarchique ou pyramidale ascendante est appliqué en utilisant un indice d'agrégation choisi $\delta$ (la dissimilarité entre les classes).

Dans HIPYR, quatre options pour $\delta$ sont disponibles :

(a) le lien simple tel que la dissimilarité entre deux classes est la dissimilarité minimum entre les membres d'une classe et les membres de l'autre classe,

(b) le lien complet tel que la dissimilarité entre deux classes est la dissimilarité maximum entre les membres d'une classe et les membres de l'autre classe,

(c) le lien moyen tel que la dissimilarité entre deux classes est la moyenne des dissimilarités entre les membres d'une classe et les membres de l'autre classe et

(d) le diamètre qui est tel que la dissimilarité entre deux classes est la dissimilarité maximum entre les membres des deux classes ensemble.

Les classes d'unités sont construites de manière récursive, c'est-à-dire qu'à chaque pas, une nouvelle classe $C$ est formée par la fusion de classes appropriées $C_\alpha$ et $C_\beta$ construites précédemment, selon la structure de classification appropriée :

- si la structure est une hiérarchie : aucune des deux classes n'a encore été regroupée ;
- si la structure est une pyramide : aucune des deux classes n'a encore été regroupée deux fois, et il existe un ordre sur $E$ tel que toutes les classes formées précédemment, ainsi que C, sont des intervalles de cet ordre.

Notons $P_t$ l'ensemble des classes formées après le pas $t$ et $S_t \subseteq P_t \times P_t$ l'ensemble des paires d'éléments de $P_t$ pouvant être regroupés ensemble au pas $t+1$, selon le modèle choisi.

Nous présentons maintenant un algorithme général de la classification classique :

- Initialisation :
  $C_i = \{w_i\}$ et $f(C_i) = 0$, $\forall i = 1, \ldots, n$
  $P_0 = \{\{w_i\},\ \forall i = 1, \ldots, n\}$
  $S_0 = P_0 \times P_0$
- Agrégation :
  **Après le pas $t$, on a :**
  $P_t = \{C_h, h = 1, \ldots, m\}$
  $S_t = \{(C_h, C_{h'}) \subseteq P_t \times P_t : C_h$ peut être groupée avec $C_{h'}\}$
  **Tant que** $E \neq P_t$ **:**
  Soit $(\alpha, \beta)$ : $\delta(C_\alpha, C_\beta) = \min\{\delta(C_h, C_{h'}) \mid (C_h, C_{h'}) \in S_t\}$
  Alors $C_{m+1} = C_\alpha \cup C_\beta$
  $f(C_{m+1}) = \max\{\delta(C_\alpha, C_\beta), f(C_\alpha), f(C_\beta)\}$
  $P_{t+1} = P_t \cup \{C_{m+1}\}$

### 6.5.2 Classification symbolique

La classification symbolique signifie que les hiérarchies ou les pyramides sont construites sur base d'un ensemble de données symboliques. Dès lors, l'algorithme présente certaines caractéristiques différentes du précédent. En effet, chaque classe est une paire $(C, s)$, appelée concept, qui est décrite "en extension" par ses membres et "en intention" par leur description. Chaque classe a une représentation symbolique automatique grâce à un objet symbolique.

#### Généralisation du représentant symbolique

Soient deux classes $C$ et $C'$ telles que $C \subseteq C'$, notons $s'$ le représentant de $C'$ et $s$ le représentant de $C$. Alors on dit que $s$ est plus général que $s'$ si son extension contient l'extension de $s'$, c'est-à-dire que $s'$ est plus spécifique que $s$.

La généralisation de deux objets symboliques $s$ et $s'$, déterminant $s''$, implique que $s''$ est plus général que les deux $s$ et $s'$.

$$s \leq s \cup s' \qquad \text{et} \qquad s' \leq s \cup s'$$

$$ext(s) \subseteq ext(s \cup s') \qquad \text{et} \qquad ext(s') \subseteq ext(s \cup s') \ .$$

La procédure de généralisation diffère selon le type de variable : intervalle, catégorique multivaluée, modale. Les trois cas sont décrits ci-dessous.

1. **Variables intervalles :**
   Soit $C_1$ une classe dont le représentant est

   $$s_1 = [y \in [a_1, b_1]]$$

   et $C_2$ une autre classe telle que son représentant est

   $$s_2 = [y \in [a_2, b_2]] \ .$$

   Si on regroupe ces deux classes pour obtenir $C$, alors son représentant $s$ est tel que

   $$s = s_1 \cup s_2 = [y \in [\min\{a_1, a_2\}, \max\{b_1, b_2\}]] \ .$$

   **Exemple 6.5.1**
   Supposons que la variable intervalle représente le temps passé à lire son journal.
   Soit

   $$s_1 = [\text{temps} \in [5, 15]] \ ,$$
   $$s_2 = [\text{temps} \in [10, 20]] \ ,$$

   alors

   $$s_1 \cup s_2 = [\text{temps} \in [5, 20]] \ .$$

2. **Variables catégoriques multivaluées :**
   Soit $C_1$ une classe dont le représentant est

$$s_1 = [y \in V_1]$$

   et $C_2$ une autre classe telle que son représentant est

$$s_2 = [y \in V_2] \ .$$

   Si on regroupe ces deux classes pour obtenir $C$, alors son représentant $s$ est tel que

$$s = s_1 \cup s_2 = [y \in V_1 \cup V_2] \ .$$

   **Exemple 6.5.2**
   Supposons que la variable catégorique multivaluée représente les professions existantes dans une société.
   Soit

$$s_1 = [\text{profession} \in \{ \text{ secrétaire, enseignant } \}] \ ,$$

$$s_2 = [\text{profession} \in \{ \text{ employé } \}] \ ,$$

   alors

$$s_1 \cup s_2 = [\text{profession} \in \{ \text{ secrétaire, enseignant, employé } \}] \ .$$

3. **Variables modales :**
   Deux possibilités sont proposées :
   $\longrightarrow$ prendre pour chaque catégorie le **Maximum** de ses fréquences
   $\longrightarrow$ prendre pour chaque catégorie le **Minimum** de ses fréquences

   Soit $\{m_1, \ldots, m_k\}$ l'ensemble des catégories d'une variable $y$ et notons les fréquences des catégories $p_j$ telles que

$$0 \leq p_j \leq 1 \ , j = 1, \ldots, k \qquad (p_1 + \ldots + p_k = 1) \ .$$

   Soit $C_1$ une classe dont les fréquences de distribution sur des variables discrètes sont

$$s_1 = [y \in \{m_1(p_1^1), \ldots, m_k(p_k^1)\}]$$

et $C_2$ telle que

$$s_2 = [y \in \{m_1(p_1^2), \ldots, m_k(p_k^2)\}] \ .$$

(a) Généralisation par le Maximum :

Lorsqu'on regroupe ces deux classes en une seule classe $C$, la généralisation de l'objet symbolique par le Maximum consiste à prendre les maximums des fréquences de la manière suivante.

$$s_1 \cup s_2 = [y \in \{m_1(p_1^1), \ldots, m_k(p_k^1)\}] \cup [y \in \{m_1(p_1^2), \ldots, m_k(p_k^2)\}]$$

$$= [y = \{m_1(p_1), \ldots, m_k(p_k)\}] = s$$

avec $p_j = \max\{p_j^1, p_j^2\}$ .

L'extension de $s$ est alors telle que

$$ext(s) = \{a : p_j^a \leq p_j, \ j = 1, \ldots, k\} \ ,$$

cela s'appelle le principe "au plus".

**Exemple 6.5.3**
Supposons que la variable modale représente les proportions d'emplois occupés par les différentes professions existantes dans une entreprise.

Soit

$s_1$ = [Type de profession $\in$ {(0.3) administration, (0.7) enseignement}] ,

$s_2$ = [Type de profession $\in$ {(0.6) admin., (0.2) enseig., (0.2) secrétariat}] ,

alors le nouveau représentant est le suivant :

$s_1 \cup s_2$ = [Type de profession $\in$ {(0.6) admin., (0.7) enseig., (0.2) secrét.}] .

(b) Généralisation par le Minimum :

Lorsqu'on regroupe ces deux classes en une seule classe $C$, la généralisation de l'objet symbolique par le Minimum consiste à prendre les minimums des fréquences de la manière suivante.

$$s_1 \cup s_2 = [y \in \{m_1(p_1^1), \ldots, m_k(p_k^1)\}] \cup [y \in \{m_1(p_1^2), \ldots, m_k(p_k^2)\}]$$

$$= [y \in \{m_1(p_1), \ldots, m_k(p_k)\}] = s$$

$$\text{avec } p_j = \min\{p_j^1, p_j^2\} \ .$$

L'extension de $s$ est alors telle que

$$ext(s) = \{a : p_j^a \geq p_j, \ j = 1, \ldots, k\} \ ,$$

cela s'appelle le principe "au moins".

**Exemple 6.5.4**
En reprenant l'exemple précédent (6.5.3), on obtient comme nouveau représentant, la solution suivante :

$$s_1 \cup s_2 = [\text{Type de profession} \in \{(0.3) \text{ admin.}, (0.2) \text{ enseig.}\}] \ .$$

Maintenant une description exacte de la manière dont on trouve le degré de généralité est donnée au point suivant.

### Mesure du degré de généralité

Dans le cas de variables symboliques on peut décrire la classe $C$ comme étant l'ensemble de tous les éléments dont la valeur des variables est acceptée par son objet symbolique $s$.

Plus formellement, pour chaque variable $i$, notons $O_i$ l'espace borné de toutes les modalités prises par cette variable et $V_i$ l'espace des modalités couvert par l'objet symbolique $s$ pour la variable $i$. On peut alors noter

$$C = \wedge \ [y_i \in V_i] \qquad \text{et} \qquad V_i \subseteq O_i \ .$$

En considérant que la classification se fait sur base de $i$ variables symboliques, la régle générale du calcul du degré de généralité peut s'écrire de la manière suivante :

$$G(s) = \prod_{i=1}^{p} \frac{m(V_i)}{m(O_i)} = \prod_{i=1}^{p} G(e_i) \qquad \text{si } s = \wedge\, e_i \ .$$

Le degré de généralité $G(s)$ représente donc la proportion d'espace couvert par la classe $C$.

Selon que les variables soient intervalles ou catégoriques multivaluées, les mesures $m(V_i)$ et $m(O_i)$ sont particulières, et pour les variables modales, le calcul de $G(s)$ est un peu plus complexe.

1. **Pour les variables intervalles :**
   Dans ce cas-ci, la mesure de l'étendue d'espace couverte est telle que

   $$m(V_i) = \max V_i - \min V_i \ .$$

   **Exemple 6.5.5**
   Considérons la description de groupes de personnes sur lesquelles ont été définies les variables "âge" et "salaire". La variable "âge" varie de la valeur 15 à 60, le variable "salaire" de 0 à 10000.
   Prenons un groupe décrit par l'objet symbolique

   $$s_1 = [\text{âge } \in [20, 45]] \wedge [\text{ salaire } \in [1000, 3000]] = e_{11} \wedge e_{12} \ .$$

   Alors on obtient

   $$G(e_{11}) = \frac{45 - 20}{60 - 15} = \frac{25}{45} = 0.55 \qquad \text{et}$$

   $$G(e_{12}) = \frac{3000 - 1000}{10000 - 0} = \frac{2000}{10000} = 0.2$$

   et la solution du degré de généralité de l'objet symbolique représentant le groupe est la suivante :

   $$G(s_1) = 0.55 \times 0.2 = 0.11 \ .$$

2. **Pour les variables catégoriques multivaluées :**
   La mesure de la proportion d'espace couvert se calcul de la manière suivante,

$$m(V_i) = \#V_i \ .$$

   **Exemple 6.5.6**
   Considérons la description de groupes de personnes de l'UE par les variables "sexe" et "nationalité".
   Prenons, en particulier, un groupe décrit par l'objet symbolique :

$$s_1 = [\text{sexe} \in \{M\}] \wedge [\text{nationalité} \in \{\text{Français Anglais}\}] = e_{11} \wedge e_{12} \ .$$

   On obtient

$$G(e_{11}) = \frac{1}{2} = 0.5 \qquad \text{et}$$

$$G(e_{12}) = \frac{2}{25} = 0.08.$$

   Dès lors le degré de généralité de l'objet symbolique du groupe vaut

$$G(s_1) = 0.5 \times 0.08 = 0.04 \ .$$

3. **Pour les variables modales :**

   (a) Généralisation par le Maximum :

   La mesure du degré de généralité d'un objet décrit par les modalités $p_i \forall i = 1, \ldots, k$ se calcule de la manière suivante dans le cas où les $p_i$ ont été obtenues avec la généralisation par le maximum,

$$G_1(a) = \prod_{j=1}^{p} \frac{1}{\sqrt{k_j}} \sum_{i=1}^{k_j} \sqrt{p_{ij}} \ ,$$

   qui est le *coefficient d'affinité*, (Voir référence [9]), entre $(p_1, \ldots, p_k)$ et la distribution uniforme.

   $G_1(a)$ est maximum (=1) quand $p_i = 1/k, \forall i = 1, \ldots, k$, c'est-à-dire une distribution uniforme.
   Donc au plus la distribution d'un objet ressemble à la distribution uniforme et au plus cet objet est général.

(b) Généralisation par le Minimum :

Dans le cas où les $p_i$ ont été obtenues avec la généralisation par le minimum, on a que

$$G_2(a) = \prod_{j=1}^{p} \frac{1}{\sqrt{k_j(k_j-1)}} \sum_{i=1}^{k_j} \sqrt{(1-p_{ij})} \ .$$

Ici encore, $G_2(a)$ est maximum (=1) quand $p_i = 1/k, \forall i = 1, \ldots, k$, c'est-à-dire une distribution uniforme.

### Méthode générale et algorithme

Soit $E = \{w_1, \ldots, w_n\}$ l'ensemble des unités que nous souhaitons classer, et $s_i$ l'objet symbolique associé à $w_i, \forall i = 1, \ldots, n$. Il est supposé que tous les $(w_i, s_i), \forall i = 1, \ldots, n$ sont des concepts et l'ensemble initial des concepts est justement $\{(w_1, s_1), \ldots, (w_n, s_n)\}$.

Soit $s = s_\alpha \cup s_\beta$ et $G(s)$ le degré de généralité de $s$, alors les classes $C_\alpha$ et $C_\beta$ doivent remplir les conditions suivantes pour être groupées en une nouvelle classe $C$ représentée par $s$ :

**a.** $C_\alpha$ et $C_\beta$ peuvent être agrégées ensemble selon la structure de classification désirée.

**b.** $s$ est complet et $s$ est plus général que $s_\alpha$ et $s_\beta$.

**c.** $ext_E(s) = C$, en d'autres termes, aucun élément de $E$ en dehors de $C$ n'appartient à l'extension de $s$, c'est-à-dire ne remplit pas les conditions exprimées par $s$.

**d.** Un critère numérique est minimum, ce critère peut être le degré de généralité de l'objet symbolique résultant $s$, $G(s)$ ou l'augmentation du degré de généralité.

Alors le concept correspondant à la nouvelle classe est $(C, s)$. Si aucune paire de classes $C_\alpha, C_\beta$ ne remplit les conditions **a.** et **b.**, l'algorithme procède en essayant de grouper plus que deux classes à la fois (avec une adaptation convenable des conditions de regroupement).

En utilisant le critère du degré de généralité minimum pour choisir parmi les paires de classes $C_\alpha, C_\beta$ remplissant les conditions **a.** et **b.**, l'algorithme

forme en premier des classes qui sont associées aux objets symboliques les moins généraux.

La nouvelle classe $C$ est indicée par $f(C) = G(s)$, la valeur du degré de généralité de $s$. L'algorithme touche à sa fin quand la classe $E$ est formée, laquelle correspond à un concept $(E, s)$ pour un $s$ convenable.

L'algorithme suivant construit une hiérarchie ou une pyramide indicée au sens large, telle que chaque classe formée correspond à un concept.

Notons encore $P_t$ l'ensemble des classes formées après le pas $t$, notons en plus $Q_t$ l'ensemble des concepts et $S_t \subseteq P_t \times P_t$ l'ensemble des paires d'éléments de $P_t$ qui peuvent être agrégées au pas $t+1$, selon le modèle choisi. Pour simplifier la présentation de l'algorithme, nous supposerons que $S_t \neq \emptyset$ à tous les pas.

- Initialisation :

  $C_i = \{w_i\}$ et $f(C_i) = 0$, $\forall i = 1, \ldots, n$ ,
  $P_0 = \{\{w_i\},\ \forall i = 1, \ldots, n\}$ ,
  $Q_0 = \{(w_1, s_1), \ldots, (w_n, s_n)\}$,
  $S_0 = P_0 \times P_0$.

- Agrégation :

  **Après le pas $t$, on a :**
  $P_t = \{C_h, h = 1, \ldots, m\}$,
  $Q_t = \{(C_h, s_h), h = 1, \ldots, m\}$
  $S_t = \{(C_h, C_{h'}) \subseteq P_t \times P_t : C_h$ peut être groupée avec $C_{h'}\}$.

  **Tant que** $E \neq P_t$ **:**
  $(\star)$ Soit $(\alpha, \beta)$ : $G(s_\alpha \cup s_\beta) = \min\{G(s_h \cup s_{h'}) \mid (C_h, C_{h'}) \in S_t\}$
  **Si** $ext_E(s_\alpha \cup s_\beta) = C_\alpha \cup C_\beta$
  **Alors**
  $C_{m+1} = C_\alpha \cup C_\beta$
  $s_{m+1} = s_\alpha \cup s_\beta$
  $f(C_{m+1}) = G(s_\alpha \cup s_\beta)$
  $P_{t+1} = P_t \cup \{C_{m+1}\}$
  $Q_{t+1} = Q_t \cup \{(C_{m+1}, s_{m+1})\}$
  **Sinon**
  $S_t = S_t \setminus (C_\alpha, C_\beta)$
  **Retourner en** $(\star)$

Remarquons qu'il est également possible d'utiliser l'augmentation minimale de généralité comme critère d'agrégation des classes, à la place du minimum absolu du degré de généralité.

### 6.5.3 Le choix des paramètres

Maintenant que les critères d'agrégation possibles ont été décrit plus clairement, qu'il a été dit que l'on pouvait choisir une classification classique sur base d'une matrice de dissimilarité ou une classification symbolique sur base d'objets symboliques, montrons la fenêtre des paramètres, figure 6.2, proposée sur SODAS 2 avant de pouvoir lancer le module HIPYR. On peut premièrement choisir de suivre soit le modèle hiérarchique, soit le modèle pyramidale. Ensuite, il faut dire si on veut une méthode classique ou symbolique et il faut choisir entre les deux critères, soit le degré de généralité minimal, soit l'augmentation du degré de généralité minimale. On peut également obtenir la matrice induite si on le désire. Enfin, le module HIPYR peut être lancé une fois que tous les paramètres sont bien fixés.

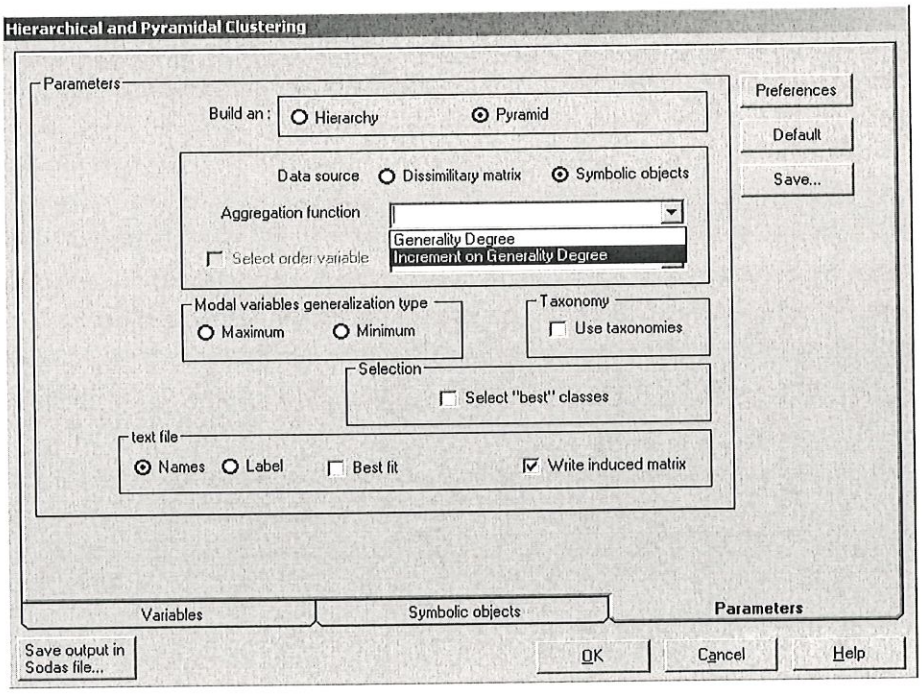



FIG. 6.2 – Fenêtre des paramètres pour le module HIPYR

# Chapitre 7

# Applications

Bien que HIPYR soit capable de fournir aussi bien des hiérarchies que des pyramides, les applications traitées dans ce chapitre concerneront uniquement les pyramides car c'est le sujet concerné dans ce travail.

## 7.1 La base de données *artificiel.sds*

Une première étude de classification pyramidale a été faite sur la base de données *artificiel.sds*. Cette base contient une description de 10 huiles différentes décrites par deux variables, "specific" dont les valeurs s'étendent de 0.5 à 12 et "freezing" qui varie de -6 à 6.

Comme il est intéressant de comparer la classification pyramidale avec le critère du degré de généralité minimal et la classification avec le critère de l'augmentation du degré de généralité minimale, ces comparaisons vont être faites pour cette application.

Tout d'abord, on va commencer par la classification avec le critère du degré de généralité minimal en montrant les différents résultats et la pyramide obtenue, ensuite on fera la même chose pour la classification avec le critère de l'augmentation du degré minimale et enfin, on comparera les résultats obtenus avec les deux méthodes.

### 7.1.1 Critère du degré de généralité minimum

Pour une analyse relativement complète des caractéristiques des différentes classes obtenues, le tableau suivant donne la hauteur des paliers ainsi

que la valeur des objets symboliques des classes. Par contre pour connaître quels sont les éléments appartenant aux classes, il faut se reporter au graphique de la figure 7.1 où la construction des paliers est bien distincte.

Dans la section 6.5.2, il avait été dit qu'on attribuait aux paliers une hauteur égale au degré de généralité, mais en fait pour plus de clarté dans la représentation graphique les hauteurs d'agrégation données dans le listing de sortie sont les degrés de généralité multipliés par 2.

| Classe | Hauteur d'agrégation | Objet symbolique |
|---|---|---|
| 1 | 0.0869565 | s=[specific = [8, 10]] $\wedge$ [freezing = [1, 4]] |
| 2 | 0.0869565 | s=[specific = [10.5, 12]] $\wedge$ [freezing = [2, 6]] |
| 3 | 0.115942 | s=[specific = [8, 12]] $\wedge$ [freezing = [2, 4]] |
| 4 | 0.126812 | s=[specific = [3, 5.5]] $\wedge$ [freezing = [2.5, 6]] |
| 5 | 0.144928 | s=[specific = [3, 4]] $\wedge$ [freezing = [-6, 4]] |
| 6 | 0.173913 | s=[specific = [8, 12]] $\wedge$ [freezing = [1, 4]] |
| 7 | 0.181159 | s=[specific = [0.5, 5.5]] $\wedge$ [freezing = [3.5, 6]] |
| 8 | 0.231884 | s=[specific = [8, 12]] $\wedge$ [freezing = [2, 6]] |
| 9 | 0.253623 | s=[specific = [0.5, 5.5]] $\wedge$ [freezing = [2.5, 6]] |
| 10 | 0.289855 | s=[specific = [8, 12]] $\wedge$ [freezing = [1, 6]] |
| 11 | 0.333333 | s=[specific = [0.5, 12]] $\wedge$ [freezing = [4, 6]] |
| 12 | 0.416667 | s=[specific = [0.5, 12]] $\wedge$ [freezing = [3.5, 6]] |
| 13 | 0.434783 | s=[specific = [3, 5.5]] $\wedge$ [freezing = [-6, 6]] |
| 14 | 0.869565 | s=[specific = [0.5, 5.5]] $\wedge$ [freezing = [-6, 6]] |
| 15 | 2 | s=[specific = [0.5, 12]] $\wedge$ [freezing = [-6, 6]] |

Suite à ce tableau, on peut déjà remarquer que les paliers des deux premières classes ont la même hauteur. Cependant les objets symboliques représentant ces deux classes sont différents.

Apparemment, lorsqu'il faut choisir entre deux classes ayant le même degré de généralité, l'algorithme agrège d'abord la classe dont l'objet symbolique intervalle a une valeur minimum plus petite que celle de l'objet symbolique de l'autre classe. C'est-à-dire que par exemple, pour les classes 1 et 2, l'algorithme construit d'abord la classe 1 car les valeurs minimales des

intervalles sont 8 et 1 tandis que pour la classe 2 les valeurs minimales des intervalles sont 10.5 et 2, comme $8 < 10.5$ et $1 < 2$, la classe 1 est agrégée en premier et puis on passe à la classe 2.

On a, comme prévu, la hauteur de la dernière classe qui vaut 1 car c'est en fait le degré de généralité de cette classe et la généralité doit être maximum car c'est la dernière à être construite et tous les éléments sont contenus dedans.

Regardons maintenant la pyramide obtenue, elle fournit des informations supplémentaires telles que l'appartenance des éléments aux classes.

Prenons par exemple la classe 6 qui est l'union des classes 1 et 3, en reprenant les objets symboliques de ces deux classes on voit clairement que l'objet symbolique de la classe 6 est bien l'union des objets symboliques des deux classes 1 et 3, on a bien $[8, 12] = [8, 10] \cup [8, 12]$ et $[1, 4] = [1, 4] \cup [2, 4]$.

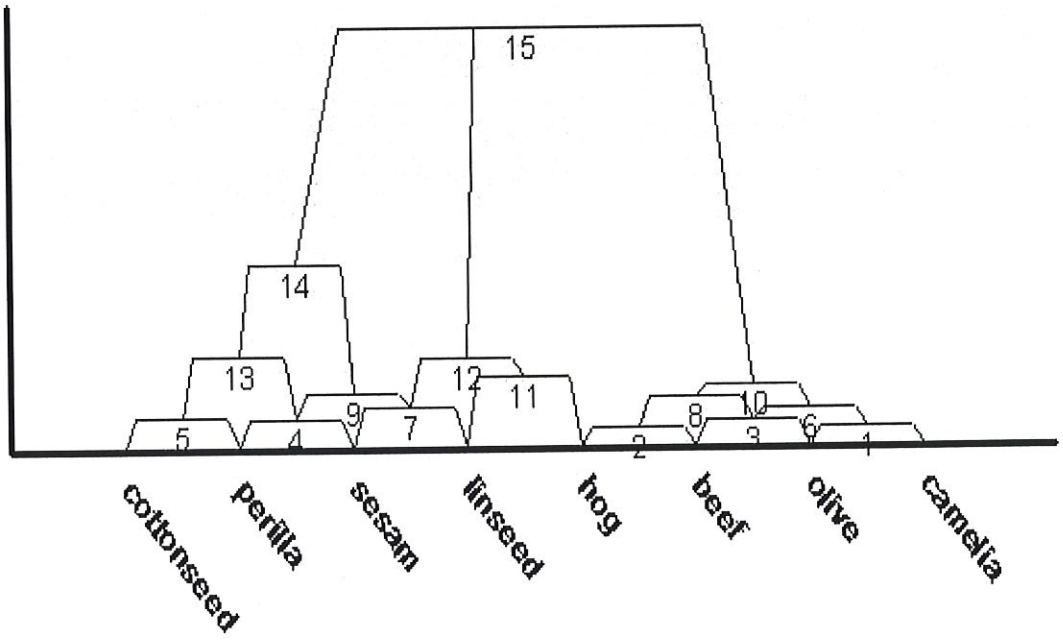



FIG. 7.1 – Pyramide obtenue avec le critère du degré de généralité minimal

Afin de ne pas répéter la même chose pour la description de chacune des classes, laissons le tableau des données et la pyramide parler d'eux-mêmes car toutes les informations y sont contenues.

### 7.1.2 Critère de l'augmentation du degré de généralité minimale

De la même manière que pour le point précédent, on dispose du tableau fournissant les objets symboliques et les hauteurs des classes ainsi que de la pyramide créée par VPYR.

| Classe | Hauteur d'agrégation | Objet symbolique |
|---|---|---|
| 1 | 0.0869565 | [specific = [8, 10]] ∧ [freezing = [1, 4]] |
| 2 | 0.0869565 | [specific = [10.5, 12]] ∧ [freezing = [2, 6]] |
| 3 | 0.115942 | [specific = [8, 12]] ∧ [freezing = [2, 4]] |
| 4 | 0.173913 | [specific = [8, 12]] ∧ [freezing = [1, 4]] |
| 5 | 0.126812 | [specific = [3, 5.5]] ∧ [freezing = [2.5, 6]] |
| 6 | 0.144928 | [specific = [3, 4]] ∧ [freezing = [-6, 4]] |
| 7 | 0.231884 | [specific = [8, 12]] ∧ [freezing = [2, 6]] |
| 8 | 0.289855 | [specific = [8, 12]] ∧ [freezing = [1, 6]] |
| 9 | 0.181159 | [specific = [0.5, 5.5]] ∧ [freezing = [3.5, 6]] |
| 10 | 0.253623 | [specific = [0.5, 5.5]] ∧ [freezing = [2.5, 6]] |
| 11 | 0.434783 | [specific = [3, 5.5]] ∧ [freezing = [-6, 6]] |
| 12 | 0.333333 | [specific = [0.5, 12]] ∧ [freezing = [4, 6]] |
| 13 | 0.416667 | [specific = [0.5, 12]] ∧ [freezing = [3.5, 6]] |
| 14 | 0.869565 | [specific = [0.5, 5.5]] ∧ [freezing = [-6, 6]] |
| 15 | 2 | [specific = [0.5, 12]] ∧ [freezing = [-6, 6]] |

Ici encore, on voit bien qu'une classe qui est l'union de deux autres classes a bien comme objet symbolique l'union des objets symboliques des deux classes.

Prenons un exemple avec la classe 14 qui est formée par le regoupement des classes 10 et 11 (voir le graphique de la figure 7.2). On a bien que son

objet symbolique, [specific = [0.5, 5.5]] ∧ [freezing = [-6, 6]], est l'union des objets symboliques des classes 10 et 11.

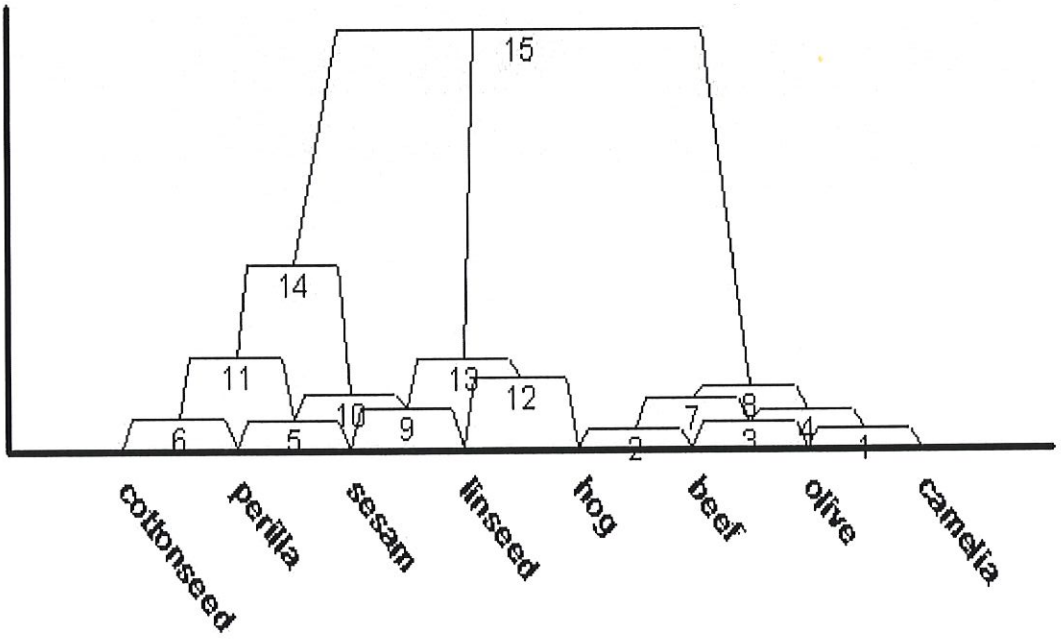



FIG. 7.2 – Pyramide obtenue avec le critère d'augmentation du degré de généralité minimale

### 7.1.3 Comparaisons des deux méthodes

Il est évident que les deux pyramides précédentes sont très semblables mais il y a tout de même une différence importante entre elles, il s'agit de l'indiçage des paliers. Bien que l'ensemble des paliers soit identique pour les deux pyramides, on remarque que les paliers n'ont pas toujours été créés au même moment. Par exemple, le palier numéroté 4 n'est pas le même dans les deux graphiques, le palier 4 de la figure 7.2 correspond au palier 6 du graphique de la figure 7.1.

Pour les deux méthodes, les classes 1, 2 et 3 sont identiques mais dans

le cas du critère de l'augmentation minimale, les deux classes 1 et 3 sont agrégées plus tôt. En effet, leur union donne lieu à la classe 4 tandis qu'avec le critère du degré minimal cette union forme la classe 6. Dans ce cas-ci, on distingue clairement la différence entre les deux critères, bien que le degré de généralité de l'union des classes 1 et 3 ne soit pas minimal, c'est cette union qui donne lieu à une augmentation de généralité minimale. On peut considérer cette mesure d'augmentation comme une mesure de distance dans le sens où lorsqu'on utilise une distance d'agrégation dans la classification classique, on ne cherche pas à agréger d'abord les plus petites classes ensemble mais bien celles qui sont les plus proches même si elles sont grandes.

Comme les ordres d'agrégation des deux méthodes commencent a être différents à partir de la classe 4, les classes qui suivent ne sont plus les mêmes.

Développons particulièrement le cas des paliers 9 et 10 de la figure 7.1 et des paliers 10 et 8 de la figure 7.2. Dans chacune des pyramides ces deux classes correspondent, la première à l'union des éléments *perilla*, *sesam* et *linseed* et la deuxième à l'union de *hog*, *beef*, *olive* et *camelia*.

Le premier graphique étant obtenu par la méthode du critère de degré de généralité minimal, on obtient d'abord la classe 9 au lieu de la 10 car les degrés de généralité valent chacun :

$$G(s_9) = \frac{5.5 - 0.5}{12 - 0.5} \times \frac{6 - 2.5}{6 + 6} = \frac{5}{11.5} \times \frac{3.5}{12} = 0.126811594 = \frac{0.253623}{2}$$

et

$$G(s_{10}) = \frac{12 - 8}{11.5} \times \frac{6 - 1}{12} = 0.144927536 = \frac{0.289855}{2} .$$

On a bien que $G(s_{10}) > G(s_9)$, donc on agrége d'abord la classe 9.

Par contre dans le cas de la deuxième méthode, la classe 8 va se former avant la classe 10 car le regroupement des classes 4 et 7 donnera lieu a une moins grande augmentation de généralité par rapport à l'union des classes 5 et 9, en témoigne les deux valeurs calculées ci-dessous, (notons $A(s_i)$ l'augmentation de généralité due à la formation de la classe $i$).

$$\begin{aligned} \text{Soit } A(s_8) &= (G(s_8) - G(s_4)) + (G(s_8) - G(s_7)) \\ &= (0.289855 - 0.173913) + (0.289855 - 0.231884) \\ &= 0.115942 + 0.057971 = 0.173913 \end{aligned}$$

$$\begin{aligned} \text{et } A(s_{10}) &= (G(s_{10}) - G(s_5)) + (G(s_{10}) - G(s_9)) \\ &= (0.253623 - 0.126812) + (0.253623 - 0.181159) \\ &= 0.126811 + 0.072464 = 0.199275 \ , \end{aligned}$$

alors on a bien que $A(s_8) < A(s_{10})$ donc la classe 8 est agrégée avant la classe 10.

Cet exemple peut être tout à fait intuitif rien qu'en regardant la pyramide. En effet, les classes 7 et 4 ont déjà deux éléments en commun et un seul élément différent l'une de l'autre tandis que les classes 5 et 9 n'ont qu'un élément en commun donc il semble logique que l'agrégation des classe 5 et 9 donne lieu à une augmentation de généralité proportionnellement plus grande que pour la classe 8.

Après l'analyse de cette application appuyée par plusieurs exemples, il ne reste qu'une chose à dire à son propos. Il s'agit de remarquer, plus concrètement, que les classes obtenues semblent assez logiques. On a les huiles animales, (*hog=porc*, *beef=boeuf*), qui sont classées ensemble dès le début ainsi que les huiles végétales (*olive*, *camelia*) et ensuite les huiles plus céréales sont aussi classées entre elles, (*cottonseed*, *perilla*, *sesam* et *linseed*).

## 7.2 La base de données *Ecotixicology.sds*

La base de données *Ecotixicology.sds* comprend la description de 12 éléments décrits par 13 variables intervalles et une variable catégorique simple. La classification pyramidale a été appliquée à cette base de données, seulement avec les variables intervalles, afin de donner un exemple dont les résultats sont très difficile à interpréter et en particulier parce que la pyramide donnée par VPYR n'apporte rien à cause de hauteurs de paliers beaucoup trop petites.

En effet, la réalisation graphique de la pyramide ne prévoit apparemment pas les cas où il existe beaucoup de paliers très bas. Lorsque la pyramide est créée, les hauteurs sont respectées et s'il existe des paliers très bas, on obtient une pyramide dans le même genre que celle obtenue pour la classification des données de *ecotixicology.sds*, voir la pyramide de la figure 7.3.

De plus HIPYR n'offre pas d'options pour lui dire de ne pas spécialement respecter les hauteurs quand la pyramide est dessinée et si on veut élargir une partie de la pyramide pour mieux voir la construction des paliers, alors on ne peut le faire que morceau par morceau.

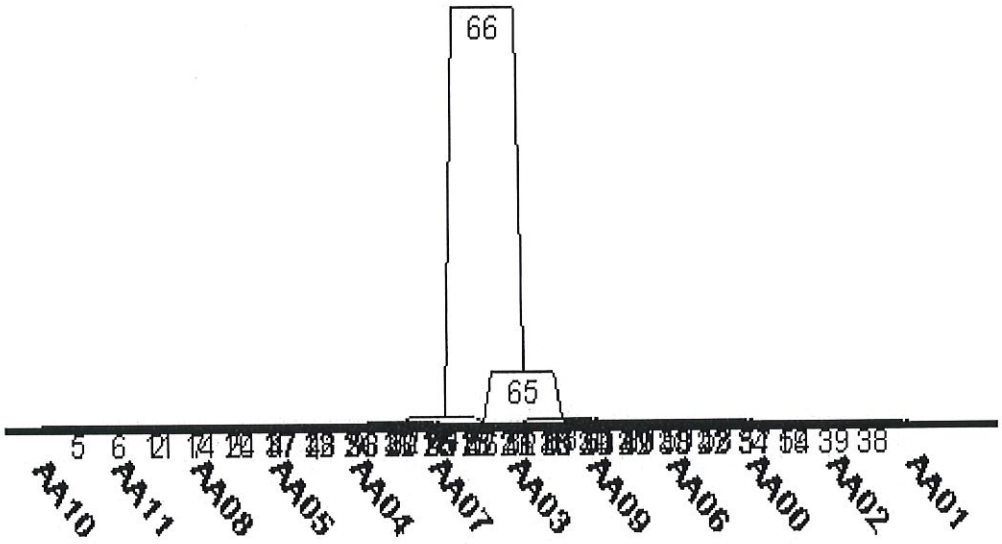



FIG. 7.3 – Pyramide obtenue avec le critère du degré de généralité minimal

Cette manière plus fine d'analyser les choses a été appliquée et a révélé qu'il n'y avait pas vraiment de structure car beaucoup de paliers sont regroupés ensemble, en même temps, dès le début. Le fait est qu'il y a quelques paliers très bas regroupant chacun beaucoup d'éléments de départ et qu'il n'y a presque pas de paliers plus haut pour terminer la pyramide.

Si on obtient des pyramides du type de la figure 7.3, on peut déjà supposer que la classification pyramidale n'a pas donné de résultats concluants quand à une structure dans les données de départ. On peut comparer cette situation au *chaining* qui se crée dans une hiérarchie, en classification hiérarchique, lorsqu'il n'y a pas de structure dans les données et que la hiérarchie regroupe successivement tous les éléments.

**Remarque :** Dans certains cas où le nombre d'éléments à classer est très grand, VPYR n'arrive pas du tout à faire la pyramide. Par exemple pour la base *temp-1974-1988.sds* qui contient 900 éléments, lorsque le programme HIPYR est lancé, ni le listing de sortie, ni le graphique produit par VPYR ne sont créés.

## 7.3 La base de données *car.sds*

Cette application a été développée dans une optique particulière, celle de l'épuration d'une pyramide saturée. Pour que l'épuration ait un intérêt, il faut travailler sur une base de données telle que la pyramide résultant de la classification pyramidale soit difficile a interpréter lorsqu'elle est saturée.

La base de données *car.sds* a été choisie car elle donne lieu à une pyramide difficile a interpréter à cause du nombre élevé de paliers. Les données représentent 33 automobiles décrites par 8 variables intervalles et le nombre de paliers obtenus pour la pyramide saturée s'élève à 419.
Une première pyramide est donnée figure 7.4 et représente la pyramide saturée obtenue après la classification avec le critère du degré de généralité minimal. Ensuite, la pyramide résultant de l'épuration, avec un paramètre valant 0.05, est donnée figure 7.5.

La pyramide 7.4 est effectivement difficile a interpréter, les paliers numérotés de 1 à environ 300 ne sont pas distinguables et on ne voit bien les recouvrements qu'à partir du moment où les paliers sont déjà numérotés dans les 300. L'épuration dans un cas comme celui-ci est presque inévitable. Nous avons, dans cette application, choisi un paramètre valant 0.05 mais on aurait pû également prendre 0.01 ou une autre valeur pas trop grande. Si le paramètre est trop élevé, on risque de supprimer trop de paliers, déjà avec $\epsilon = 0.05$, on passe de 419 paliers à 52, mais on garde cette solution. En effet la pyramide 7.5, bien que beaucoup plus allégée par rapport à la pyramide 7.4, reste toujours difficile à interpréter pour les premiers paliers donc avec $\epsilon < 0.05$ on ne simplifiera peut-être pas assez tandis qu'avec un $\epsilon > 0.05$, le nombre de paliers va certainement trop décroître, donc on travaille avec 0.05 comme paramètre d'épuration.

Un inconvénient de HIPYR réside dans le fait que lorsque l'épuration est faite, il n'est pas possible d'obtenir un nouveau fichier de sortie où on pourrait obtenir la description des classes restantes après l'épuration, on ne sait pas dire à quels paliers de départ correspondent les paliers restants. De plus, comme la base de données utilisée désigne les voitures par des labels, on ne sait pas dire, sur base de la figure 7.5, quelles voitures sont regoupées entre elles. Cependant, le but de cette application n'étant pas la classification elle-même, nous n'analyserons pas les 419 classes décrites dans le fichier de sortie.

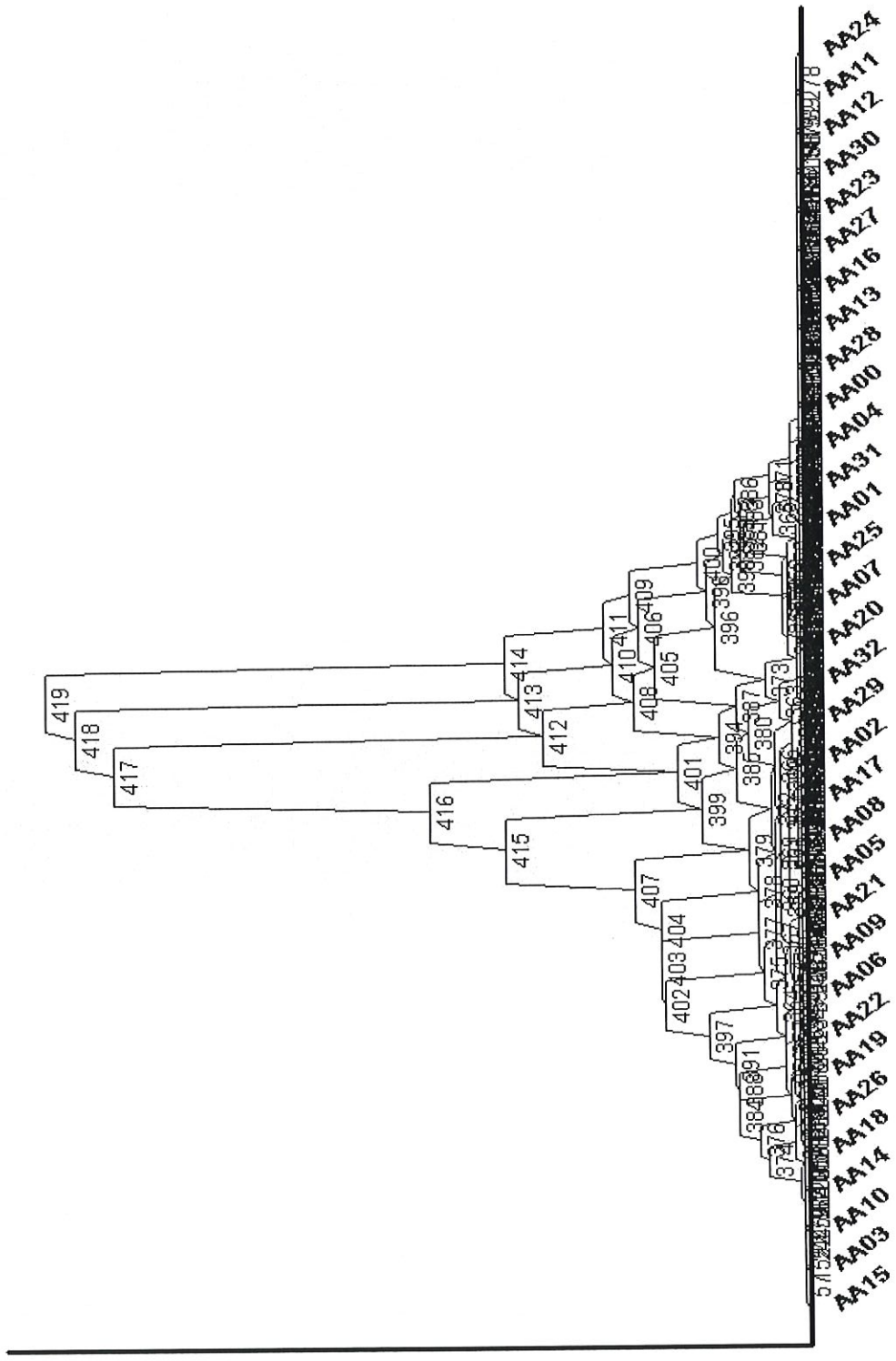



FIG. 7.4 – Pyramide saturée obtenue avec le critère du degré de généralité minimal

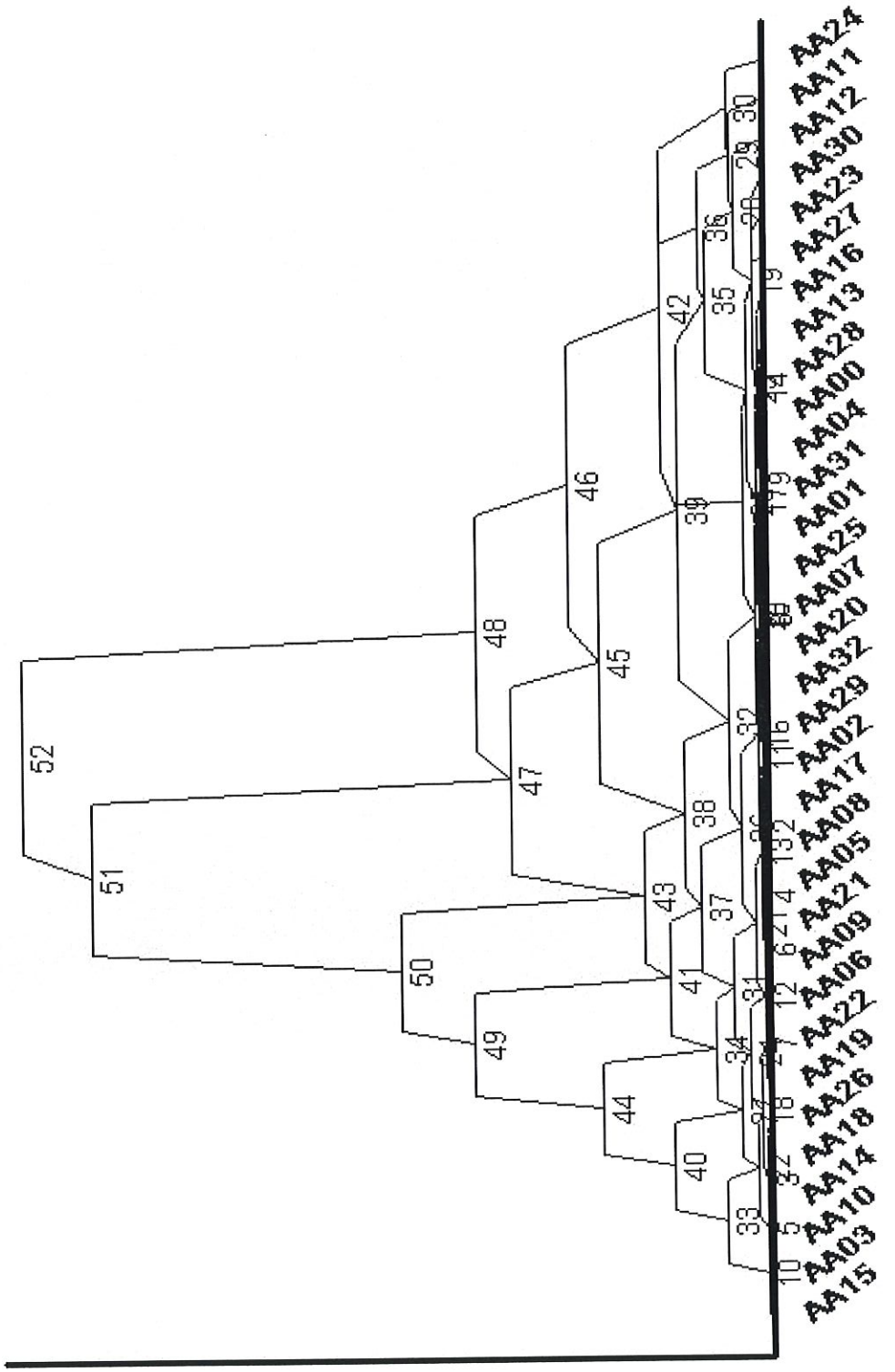



FIG. 7.5 – Pyramide épurée obtenue avec un niveau de précision de 0.05.

## 7.4 La base de données *microorganisms.sds*

Les données contenues dans la base *microorganisms.sds* représentent 10 micro-organismes décrits à l'aide de 4 variables catégoriques multivaluées.

Un dessin des éléments de cette base est donné par la figure 7.6 et ensuite une description des variables sera fournie par la table 7.1.

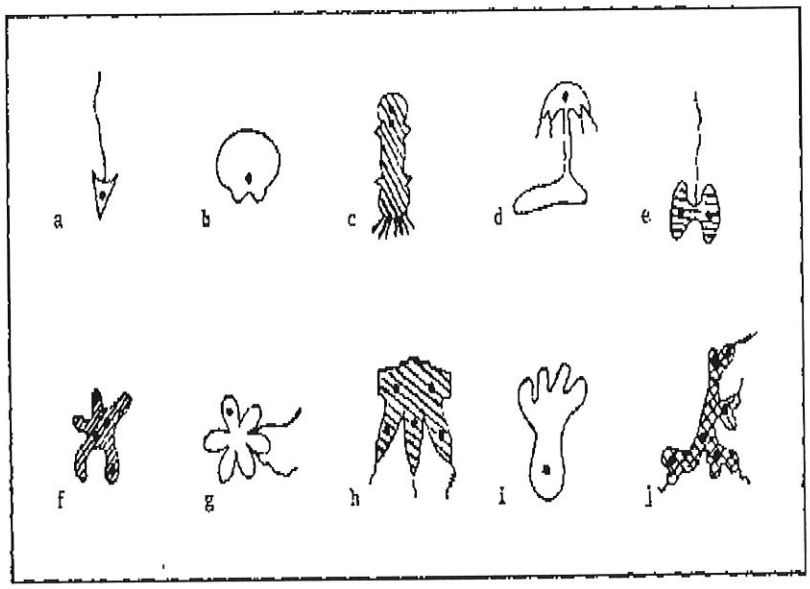



FIG. 7.6 – Les micro-organismes

La description des variables fournie par SODAS 2 est celle qui a été donnée, en 2000, par Verde, De Tenorio et Lechevallier. Les variables peuvent prendre les différentes modalités détaillées ci-dessous :

- parties du corps = (triangle, rectangle, cercle) ⇒ 3 modalités
- tâches du corps = (dessus, dessous, droite, gauche) ⇒ 4 modalités
- texture = (noir, blanc) ⇒ 2 modalités
- type de queue = (dessus, dessous, droite, aucune) ⇒ 4 modalités

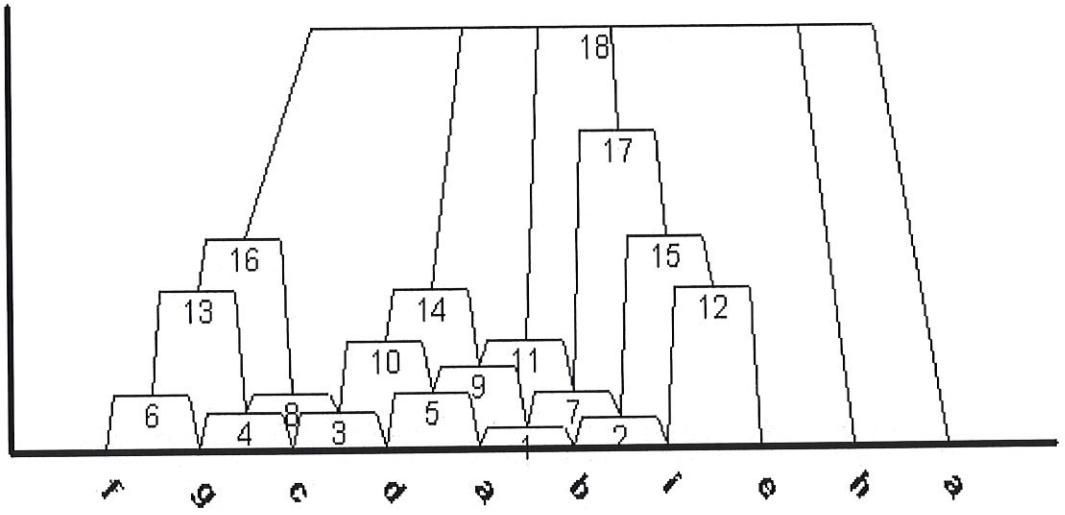



FIG. 7.7 – Pyramide obtenue avec le critère du degré de généralité minimal

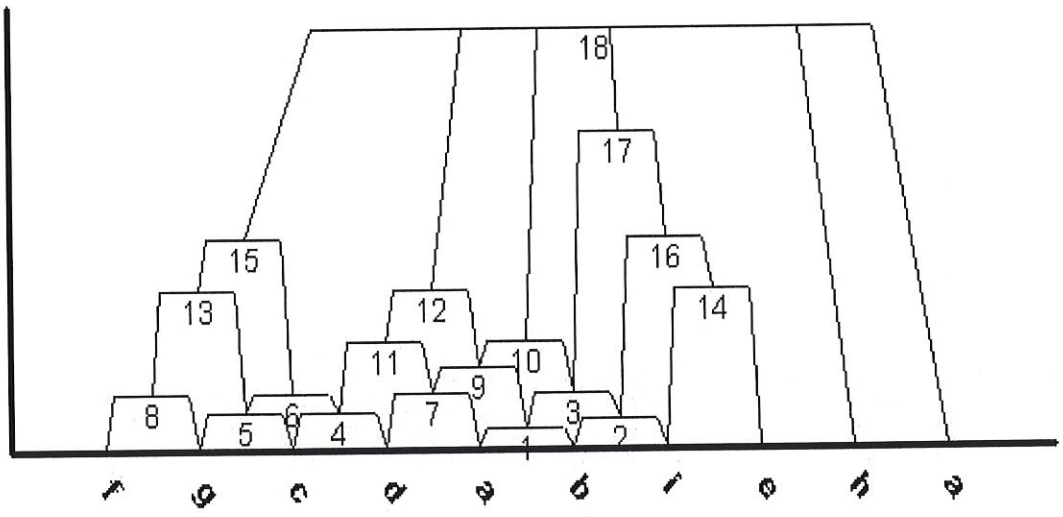



FIG. 7.8 – Pyramide obtenue avec le critère d'augmentation du degré de généralité minimale

Grâce à la simplicité des pyramides obtenues, l'analyse des résultats est assez facile à faire car on peut distinguer les différences et ressemblances rien qu'en les observant. Mais afin de présenter plus clairement les résultats obtenus et de permettre une analyse plus formelle des résultats, une liste des hauteurs des paliers obtenues pour les deux méthodes est donnée par le tableau 7.2. Ces valeurs proviennent du fichier de sortie fournit par HIPYR dans lequel sont disponibles les descriptions des classes.

| Degré de généralité minimal | | Augmentation du degré de généralité minimale | |
|---|---|---|---|
| Classe | Hauteur d'agrégation | Classe | Hauteur d'agrégation |
| 1 | 0.0416667 | 1 | 0.0416667 |
| 2 | 0.0625 | 2 | 0.0625 |
| 3 | 0.0833333 | 3 | 0.125 |
| 4 | 0.0833333 | 4 | 0.0833333 |
| 5 | 0.125 | 5 | 0.0833333 |
| 6 | 0.125 | 6 | 0.125 |
| 7 | 0.125 | 7 | 0.125 |
| 8 | 0.125 | 8 | 0.125 |
| 9 | 0.1875 | 9 | 0.1875 |
| 10 | 0.25 | 10 | 0.25 |
| 11 | 0.25 | 11 | 0.25 |
| 12 | 0.375 | 12 | 0.375 |
| 13 | 0.375 | 13 | 0.375 |
| 14 | 0.375 | 14 | 0.375 |
| 15 | 0.5 | 15 | 0.5 |
| 16 | 0.5 | 16 | 0.5 |
| 17 | 0.75 | 17 | 0.75 |
| 18 | 1 | 18 | 1 |

TAB. 7.2 – Hauteurs des paliers pour les deux méthodes

Suite à ce tableau, développons quelques exemples d'agrégation de classes pour vraiment bien comprendre comment fonctionne HIPYR

Remarquons tout d'abord que dans les deux cas, les classes 1 et 2 sont identiques. Effectivement, il semble logique que ces classes soient formées en premier puisque les éléments $a$ et $b$ ne différent que pour deux modalités (voir table 7.1) et les éléments $b$ et $i$, pour trois modalités.

Prenons l'exemple des paliers 3 et 7 de la figure 7.7 et des paliers 3 et 4 de la figure 7.8 pour comparer les différences entre les ordres de construction des paliers pour les pyramides obtenues avec les deux critères.

Dans la figure 7.8, le palier 3 est créé avant le palier 4 car son augmentation de généralité est moins grande. Par contre, observons dans le tableau 7.2 les degrés de généralité des paliers 3 et 7 de la figure 7.7 afin de montrer pourquoi le palier 3 se forme avant le palier 7 contrairement à la figure 7.8. En effet, le degré de généralité de la classe 3 vaut 0.0833333 tandis que la généralité de la classe 7 vaut 0.125, c'est donc bien la classe 3 qui doit être formée en premier car son degré de généralité est moins élevé.

Maintenant, analysons la formation des classes 10 et 11 des deux pyramides obtenues avec les deux méthodes.

Lors de la classification avec le critère d'augmentation de généralité minimale, voir figure 7.8, la raison pour laquelle la classe 10 est formée avant le palier 11 est évidente. En effet, ces deux paliers ont la même hauteur et bien que la classe 10 soit formée par des classes plus générales, (la classe 9 de hauteur 0.1875 et la classe 3 de hauteur 0.125), l'intersection de ces classes est proportionnellement plus grande que celle entre les paliers 4 et 7. En effet les classes 9 et 3 contiennent chacune trois éléments et en ont deux en commun, ceux de la classe 1. Par contre, les classes 4 et 7 contiennent chacune deux éléments et n'en ont qu'un en commun, l'élément $d$. Donc le regroupement des classes 3 et 9 donne lieu à une moins grande augmentation de généralité lors de la formation de la classe 10 que lorsque les classes 4 et 7 forment la classe 11.

Par contre, l'ordre de construction de ces deux classes est inversé dans le cas de la classification avec le critère de degré de généralité minimal, voir figure 7.7. Comme le critère d'agrégation ne repose que sur le fait que la généralité soit minimale, lorsqu'il tombe sur deux classes, 10 et 11, ayant le même degré de généralité, il prend d'abord la classe 10 formée par des classes

trouvées plus tôt, 3 et 5, et ensuite choisi la classe 11 formée par des classes obtenues plus tard, 9 et 7.

Les descriptions de cette application faites au sujet de la formation de certains paliers doivent suffire à bien comprendre comment fonctionne la classification pyramidale et la formation d'une pyramide.

"arbres épais", des "guirlandes" ou encore une courbe. Il est aussi possible de développer des problèmes d'optimisation visant à trouver la meilleure pyramide sous la forme de la recherche de l'indice pyramidal ayant la meilleure adéquation avec une distance donnée par l'utilisateur.

# Bibliographie


[1] E. DIDAY, *Une représentation visuelle des classes empiétantes : les pyramides*, Rapport de Recherche n°291, INRIA, Rocquencourt, France, 1984.

[2] E. DIDAY, *Croisements, ordres et ultramétriques*, Mathématiques des Sciences humaines, 21ème année, n°83 p. 31-54, 1983.

[3] S. DELOGNE, *Méthodes de détermination du nombre de classes pour des données symboliques de types intervalle*, Mémoire, FUNDP, Namur, 2002.

[4] H.-H. BOCK ET E. DIDAY, *Analysis of Symbolic Data*, Springer, Berlin, 2000.

[5] A. HARDY, *Aspects statistiques de la classification*, Notes de cours, FUNDP, Namur, 2003-2004.

[6] P. BRITO, *Hierarchical and Pyramidal Clustering - Clustering and Visualizing Symbolic Data by using the modules HIPYR and VPYR*, ASSO Project, Porto, 2003.

[7] P. BRITO, *Classification hiérarchique et pyramidale de données symboliques*, Notes du Séminaire donné à Namur, mai 2004.

[8] BARBUT ET MONJARDET, *Ordre et classification*, Classiques Hachette, 1970.

[9] K. MATUSITA, *Decision rules based on distance for problems of fit, two samples and estimation*, Ann. Math. Stat. 3, 1-30, 1951.